3 ummer Part 2 of 2

# Chapter 4: Are Value-Added Models an Option for Michigan?

As part of its pilot of educator effectiveness tools, MCEE asked ISR researchers to explore whether value-added modeling (VAM) might be a practical approach for the state to fulfill the "student growth" tool requirements of PA 102 of 2011. In simplest terms, value-added modeling attempts to measure a teacher's impact on student achievement (the "value" he or she adds) *apart from* other factors that influence students' achievement, such as individual ability, socio-economic factors, and peer influences. The relevance of value-added modeling to PA 102 of 2011 is considerable given data reported in Chapter 2 of this report. That chapter described the relatively unsystematic approaches to measuring student growth that were developed in pilot school systems, due in part to the lack of timely availability of state assessment data for measuring student growth. Value-added modeling (VAM) represents one way to address this issue. It would use Michigan's state assessment system to develop fair and uniform standards for measuring teachers' impacts on student achievement.

This chapter discusses ISR's work with three VAM vendors and examines the feasibility of deploying value-added measures of teaching effectiveness based on Michigan's state testing data for use in teacher evaluations.

## VAM Pilot Data and Procedures

To address this issue, ISR contracted with three vendors, each of which has had extensive experience conducting value-added analyses for state and local education agencies. The vendors were: the American Institutes for Research (AIR), Education Analytics (EA), and SAS.

***A pilot of VAM procedures was conducted between January 2013 and October, 2013.*** During that time, ISR researchers asked the Center for Education and Performance Information (CEPI) to send ISR all currently available data (described below) needed to estimate value-added models of teaching effectiveness with MEAP data. ISR then forwarded these data to vendors. Using this process:

***The three vendors were sent three years of MEAP data on all Michigan students taking any MEAP test in Fall 2009, Fall 2010, and Fall 2011.*** Vendors were then asked to use these data to develop value-added models of student gains in achievement between Fall 2010 and Fall 2011. Note that this time interval is *two years* behind the annual teacher evaluation cycle for the pilot year of 2012-2013. This lag was due to delays in access to more recent MEAP test results.

***Each vendor also was sent data from the Michigan Student Data System (MSDS) and the end-of-year Registry of Education Personnel (EOY REP) for the years 2009, 2010, and 2011.*** Importantly, the MSDS data sent to vendors included a state developed Teacher-Student Data Link (TSDL) for the school years 2010-2011 and 2011-2012, as well as data on students' social background and educational status. EOY REP data provided a list of grades and subject areas in which each registered teacher was teaching for 2009, 2010, and 2011.

***Using MEAP, TSDL and REP data, vendors sought to identify the teacher(s) who taught each tested student over the time period Fall 2010 – Fall 2011.*** This was a complex matching procedure in which vendors took each tested student's MEAP score for a given grade/subject, then obtained that student's TSDL for the 2010-2011 school year, then verified that the teacher(s) listed as teacher of record in the data were also listed as having taught in the tested curriculum area in REP data.

***Once teacher-student linkages were created, vendors engaged in "value-added" statistical modeling.*** These models created various estimates of each student's gains in achievement (in a tested area) over the period Fall 2010 – Fall 2011 and attributed some portion of these gains to the teacher(s) to whom they were linked. These apportionments are called "teacher effects" on student achievement in the remainder of this chapter.

***Importantly, over the course of the pilot: (a) each VAM vendor used slightly different business rules to***

***process data prior to value-added modeling; and (b) each VAM vendor used different statistical models to estimate teacher effects on student achievement.*** Under ordinary business circumstances, each vendor would have preferred to consult extensively with its client prior to engaging in both these steps, but ISR researchers did not want to dictate how data were to be processed or the type(s) of value-added models to be estimated by vendors. Instead, ISR researchers asked each VAM vendor to develop a set of data processing procedures and to estimate a variety of value-added statistical models that they *might* recommend to Michigan stakeholders. In the end, this approach proved fruitful, for it illustrated the various tradeoffs that Michigan legislators (or local districts) must consider as they decide whether and how to implement value-added modeling of teacher effectiveness as a tool for teacher evaluation.

## Statistical Models Used by VAM Vendors

The first issue that policy makers seeking to use value-added modeling will need to address is the type of value-added model (VAM) they want vendors to estimate. A thorough review of this complex topic is beyond the scope of this report. However, we can begin this chapter by briefly describing the two general approaches that vendors took to value-added modeling:

***One approach to VAM analysis involves estimating what has been called a "growth model." This was one approach implemented by SAS for the pilot project.*** As implemented by SAS, this approach uses data from multiple years of student testing and generally estimates the *gains* in achievement that groups of students experience over time. In the SAS approach, teacher effects are conceptualized as "deflections" that move students' realized gains upward or downward during the time period when they are taught by a particular teacher. Importantly, the SAS approach to VAM estimation has been called a "layered" model, meaning that students' gains in achievement are assumed to be affected not only by students' current teachers, but also by past teachers (whose effects on their former students' achievement are assumed to persist over time). Thus, the teacher effect on student achievement gains estimated by SAS has been adjusted for the effects of previous teachers on a given teacher's students. Moreover, all of the teacher effects estimated in the layered model are updated annually as new data are added.

***A second approach to VAM analysis involves estimating what is typically called a "covariate adjustment" model. This was the primary approach taken by AIR and EA in the pilot (and it also was used by SAS in some analyses).*** Unlike the "growth" model, the covariate adjustment model focuses, not on gains in achievement over multiple years, but rather on a student's test score at a single point in time (for convenience, let us call this single point in time the student's "current" achievement). A covariate adjustment model essentially uses a linear regression analysis to predict a student's current achievement from many covariates, including a student's past achievement levels. Conceptually, this model assumes that students whose current test scores are *above* what is predicted by the statistical model have experienced more academic growth than students whose current scores are *below* what is predicted from the statistical model.

***Importantly: (a) all of vendors can report teachers' effects on students' achievement as effects on students at a specific grade, for a specific subject, in a specific year; and (b) all vendors can combine these estimates to provide an overall estimate of a teacher's effectiveness (across multiple grades, subjects, and years).*** This is important, for a teacher might be more or less effective at different grades, for different subjects, and in different years, but an education authority might also want a summary of teacher effectiveness that combines many separate estimates into a single average effectiveness rating. The SAS "layered" model obtains teacher effect estimates by using all available data at once; the AIR and EA covariate adjustment models estimate teacher effects separately by grade, subject, and time point, and then use sophisticated procedures to average across these estimates to get a composite score for teachers.

***All of the models estimated by VAM vendors controlled for multiple prior-year test scores in predicting students' expected gains or test score.*** In fact, this procedure is what gives rise to the term "value-added" modeling. In essence, value-added models are assuming that a student's achievement gain and/or current achievement level is affected by that students' prior achievement levels. In this sense, the VAMs do not evaluate teachers based on actual gains in achievement or the actual achievement level of students. Instead, VAMs evaluate teaching effectiveness in reference to gains or achievement levels that have been "adjusted" for students' prior levels of (or gains in) achievement, usually as measured in more than one subject, and typically as measured by data from more than one prior year. This adjustment for prior test scores is important because it means that a highly effective teacher is not always one whose students experienced the highest gains in achievement (or ended the academic year with the highest achievement scores). Rather, highly effective teachers produce achievement gains or end-of-year achievement scores that are greater than would be *predicted* based on the prior levels of achievement of the students they taught. In this sense, value-added models measure whether a given teacher is more or less effective than teachers of students with similar achievement histories.

***All of the VAM vendors also estimated statistical models that controlled for characteristics of students other than prior achievement.*** This is important, for a continuing controversy in value-added modeling is whether or not to control for characteristics of students (such as students' free lunch status, ethnicity, special education status, or other characteristics) when estimating a teacher's effects on students' achievement. In the pilot, using MEAP data, the addition of these student-level covariates typically resulted in very little change in teachers' estimated effectiveness. However, many argue that since poverty and ethnicity can affect students' achievement, estimates of teacher effects on students' achievement should take these factors into account. When such variables are entered into value-added models, value added models are measuring whether a given teacher is more or less effective than teachers of students with similar achievement histories *and* other personal characteristics (like social and economic status).

***VAM vendors also examined the extent to which estimates of teacher effects were sensitive to the aggregate composition of classrooms.*** The addition of classroom-level covariates into value-added models is another controversial issue in the field of education. Some would argue that group-level properties, such as the percentage of high-poverty students in a class or the average levels of prior achievement of students in a class (or many other variables) might affect students' achievement outcomes, largely through the process of "peer effects," where a student's classmates influence that student's learning. Vendors participating in the pilot explored the extent to which inclusion of these classroom-level covariates affected VAM estimates for teachers, and as we show at a later point in this chapter, there was evidence from vendor analyses that these covariates did affect value-added estimates for teachers. A problem, however, is that researchers cannot say with certainty whether the correlation between group-level covariates and value-added measures is the result of peer effects on students or the selection of less effective teachers into various social settings. As a result, VAM vendors typically argue that *policy makers* must decide whether or not to adjust for particular group-level covariates (like percentage of high-poverty students in a classroom or the percentage of special education students in a class) since there is no clear scientific justification for or against doing so.

***Finally, VAM vendors differed in how they estimated teacher and school effects on students' achievement.*** In general, the VAMs estimated by SAS used random effects models to estimate teacher effects and did not control explicitly for school effects. EA's VAM analyses usually estimated teacher "fixed" effects and did not estimate an explicit parameter for school effects. AIR used a random effects model to estimate teacher effects that included a random school effect. In these models, AIR added 50% of the estimated random school effect to a teacher's random effect on student achievement to produce its value-added score for a teacher.[9]

## Data Processing Issues Prior to VAM Analyses

Because there are so many variants of value-added models, it is very important for educators working with VAM vendors to engage in a planning process in which decisions are made about the type(s) of VAMs to be estimated for policy purposes. However, once a modeling option has been chosen, VAM vendors will typically proceed to the implementation phase of their work. This phase involves using the test score (and other) data provided by a state to estimate teachers' effects on students' achievement using the method(s) selected for VAM analysis.

The salient feature of the implementation phase of VAM analysis is that VAM vendors must work with data systems that are inherently complex. The data used in VAM analyses usually come to vendors as many different data sets that include data on students, data on teachers, and data on links between teachers and students. In working with these data sets, VAM vendors always use a set of "business rules" that determine how to match data across data

[9] The rationale for this approach is discussed in AIR's technical report. It should be noted, however, that AIR is not necessarily committed to this approach. Instead, this is an approach that it has used to calculate value-added scores for teachers in the State of Florida. AIR used this approach in the pilot simply to demonstrate an option for dealing with school effects in VAMs.

sets and determine which students and teachers are ultimately included in any VAM analysis.

***The data processing phase of any VAM analyses is important, for a VAM analysis is only as good as the data on which it is based.*** For this reason, an important question addressed by ISR researchers was the extent to which data collected by the State of Michigan was of sufficient quality to proceed with VAM analyses. As discussed at the beginning of this chapter, the data sets sent to vendors were: (a) MEAP test scores; (b) data on student characteristics and teacher-student linkages taken from the Michigan Student Data System (MSDS); and (c) data on teacher characteristics (including course assignments) taken from the state's Registry of Education Personnel (REP).

***Overall, VAM vendors reported that the quality of MEAP test score data was sufficiently high for sophisticated VAM analyses.*** MEAP tests were judged to be reasonably aligned to state curriculum frameworks, to have acceptable psychometric properties, and to result in normally distributed test scores for specific populations of grade/subject test takers. One VAM vendor did express concerns about the use of a Fall-to-Fall testing period (as opposed to the more common Spring-to-Spring period used in other states) arguing that the Fall-to-Fall testing period might not control for selection effects in VAM data as adequately as Spring-to-Spring testing data. In addition, students at the very floor and very ceiling of the test score distribution created problems if VAM models took errors in measurement into account. Otherwise, VAM vendors experienced few problems working with MEAP's Fall-to-Fall testing data.

***However, only about 33% of classroom teachers in Michigan are teaching classes in MEAP-tested subject/grade combinations where VAM scores might be estimated.*** Therefore, not all Michigan teachers can be evaluated using state assessment data. To see this, consider the percentage of all teachers in the state who were listed in the state's 2011 end-of-year Registry of Education Personnel (REP) as teaching any MEAP-tested subject at grades 4-8. In theory, one could estimate a VAM score for these teachers (although this is an optimistic scenario). The table to the right shows the relevant data. Teachers at 3rd grade, teachers of 9th grade social studies, and teachers of 11th grade MME tested subjects are *not* shown here because these teachers lack sufficient data for calculation of VAM scores. The numbers shown in the table

### At a Glance:<br>Michigan Teachers for Whom VAMs Can be Estimated*

| Total # of Teachers in Michigan | 93,032 |
|---|---|
| Total # Teachers of a MEAP Tested Subject | |
| • Reading | 33,589 |
| • Mathematics | 33,685 |
| • Writing | 34,416 |
| • Science | 32,896 |
| • Social Studies | 32,417 |
| # Unique Teachers of MEAP Tested Subjects Grades 4-8 | 30,196 |
| % of All Teachers Who Could Have a VAM | ≈ 33% |

*** The data in this table are *estimates* based on the number of unique teacher IDs found in 2011 EOY REP data after attributing subject teaching assignment to IDs using subject coding decisions that are very similar to those used by VAM vendors. Not all personnel listed as "teachers" are included here. The data are for teachers teaching any of the subjects listed above at MEAP tested grades 4-8. The assumption is that it is possible to estimate a VAM score for these teachers. The counts listed in the table exclude categories of personnel such as teacher consultants, various professional specialties (e.g., speech therapists), and all paraprofessionals.**

### At a Glance:<br>Number of Eligible Teachers Who Had VAMs Reported for MEAP Math and Reading *

**Mathematics**

| Grade | EOY 10-11 | AIR | EA | SAS MRM |
|---|---|---|---|---|
| 4 | 7,750 | 4,320 | 4,481 | 4,351 |
| 5 | 7,364 | 4,297 | 4,397 | 4,302 |
| 6 | 4,372 | 4,139 | 4,252 | 4,005 |
| 7 | 3,294 | 2,622 | 2,986 | 1,887 |
| 8 | 3,160 | 2,648 | 2,674 | 1,907 |

**Reading**

| Grade | EOY 10-11 | AIR | EA | SAS |
|---|---|---|---|---|
| 4 | 8,284 | 4,371 | 4,559 | 4,438 |
| 5 | 7,784 | 4,459 | 4,604 | 4,482 |
| 6 | 4,025 | 4,354 | 4,574 | 4,269 |
| 7 | 2,679 | 3,048 | 3,570 | 2,418 |
| 8 | 2,600 | 2,862 | 2,999 | 2,102 |

***These are provisional counts as of 12/11/2013 and are subject to additional verification by VAM vendors and ISR. Although provisional, the numbers in the table are unlikely to change by large fractions. They therefore illustrate how data quality and data processing procedures can reduce the number of teachers for whom VAM measures can be reported. The SAS column reports number of teachers for the MRM intra-year analysis with no student covariates other than prior test scores. The EA column reports the number of teachers reported for "method 1," a covariate adjustment model with no student covariates other than prior test scores. These numbers could change with additional data processing. The AIR column reports numbers for "model A," a covariate adjustment model with no student covariates other than prior test scores. This figure is based on ISR counts, not AIR counts.**

are rough estimates based on assumptions about types of VAMs that might be estimated. They nevertheless suggest the obvious point that if policy makers in Michigan are to develop value-added measures of teaching effectiveness for *all* teachers, including teachers at grades K-3, 9-12, and teachers of "non-academic" subjects, the state testing system will have to be expanded considerably

A second point is this: ***In the normal process of working with CEPI data, a certain percentage of teachers who teach a MEAP-tested subject cannot be included in a VAM analysis for a variety of reasons, thus further reducing the number of teachers who can be evaluated using value-added measures.*** This loss of teacher cases is demonstrated in the lower table on the previous page, which shows the consequences of data processing decisions for the number for teachers on whom VAMs were calculated. To construct this table, ISR researchers took data from the EOY 2010-2011 REP data and counted the number of teachers coded as teaching in various subject area codes at a grade. ISR researchers then used the data sets sent from VAM vendors to provide information on the number of teachers with VAM scores.

One thing that is apparent from the table is the discrepancy between the numbers of teachers ISR researchers coded as teaching reading and mathematics in grades 4-8 using EOY REP 2010-2011 data versus the number of teachers for whom VAM vendors reported value-added scores. This discrepancy occurred across all VAM vendors and almost always produced a seeming loss in teachers from the population on which value-added measures *could have been* calculated to the population on which measures *actually* were calculated. ISR researchers remain unsure of the exact causes of this seeming loss of teachers across data sets. But it should *not* be seen as a reflection of errors on the part of VAM vendors. VAM vendors were completely transparent about the business processes they used to include teachers and students in their analyses, and they worked diligently to make appropriate data linkages. From the ISR perspective, the loss of cases probably results from an interaction between the quality of data submitted to VAM vendors and the unique "business rules" used by ISR and vendors for assuring responsible reporting of VAM results.

VAM vendors presented a different analysis of the case loss process. They typically begin their analyses of case loss using student test score data. Once student test score data were in hand, for example, VAM vendors searched for the student-teacher links in the MSDS for tested students, and then turned to the EOY REP data to verify that teachers who were listed in the linkage data were also reported to have taught the relevant grade/subject combination. In essence, it appears that VAM vendors work from a target population of tested students, whereas ISR considers the EOY REP data to be the target population.

Looking at the problem of missing data from the VAM vendor perspective reveals where teacher case loss occurs as VAM vendors match students to teachers and verify that teachers are teaching at the relevant grade/subject combination. At this stage of the analysis, VAM vendors report a loss of 20-30% in teacher cases (depending on vendor, subject, and grade). About half this loss results because vendors typically drop teachers from specific subject/grade reporting when they are linked to fewer than 10 students at that grade/subject combination. The remainder of the loss occurs from missing data issues, especially for vendors using a covariate adjustment model, who drop students with missing data on prior test scores.

***Vendors' data processing work highlights a significant problem in Michigan education data. In the current data system, many Michigan teachers who work at tested grades and teach tested subjects can be linked to only a very small number of students.*** For example, in AIR data, 30% of teachers had 10 or fewer students linked to them for analysis of mathematics and reading teaching effectiveness. The problem was prevalent at all grades, but was especially aggravated at grades 7 and 8, where 25% of teachers had 2 or fewer students linked to them. In the VAM reports, these teachers do not receive a VAM score (due to small number of students), causing significant case loss in VAM reporting.

***Overall, VAM vendors were cautious about the quality of teacher-student data links and recommended that Michigan investigate possible improvements to TSDL data collection.*** Each vendor was quick to acknowledge that Michigan was in the early stages of developing a teacher-student data linkage, and each vendor was transparent about the coding decisions they made in linking teachers to students for the purposes of VAM analysis. The processes used by each vendor were thorough, but two of three vendors recommended that Michigan do more to investigate the quality of TSDL data, and two of three vendors also recommended implementation of a roster verification step as part of the teacher evaluation process (a procedure discussed in more detail in the next section of this chapter). Finally, each of the vendors has well-developed methods for working with clients to improve student-teacher linkage data. EA, for example, works directly with a client's data systems personnel. AIR and SAS do the same, but AIR also has proposed to conduct training sessions for all local school systems in Michigan designed to improve and routinize the student-teacher linkage process, a step that seems important given that TSDL (and other MSDS and REP) data are locally generated.

## The Pilot Roster Project

Because the validity of any value-added measure of a teacher's effectiveness rests crucially on identifying the subject/grade combinations taught by a teacher and identifying the students a teacher taught over a school year in those subject/grade combinations, the Michigan Council for Educator Effectiveness asked ISR researchers to conduct a pilot project to test a method for gathering accurate data on courses taught by teachers and students enrolled in those courses. In this report, this activity is called the "roster pilot."

- ***The goal of the roster pilot was to produce for each participating teacher an accurate list of courses taught by that teacher and the students enrolled in those courses over an entire school year and in doing so, to provide a convenient way for teachers to verify that roster data were accurate.***

The roster pilot's main advance was to develop a web-based interface that allowed teachers and principals to verify rosters (a major feature of SAS's implementation of VAM analysis in states where it works). The use of this interface began with ISR operations personnel requesting each district to send "roster" data much like that reported to the state to ISR for processing. ISR then used these locally-provided data to construct a list of all the classes a given teacher taught in a semester (or trimester), and for each class, all of the students the local education agency listed as being enrolled in that class.

The initial data exchanges between ISR and LEAs proved difficult. One problem was that districts in Michigan used a variety of software for student reporting; another was that district capacity for making data exchanges was quite variable; yet another was that schools varied in the way their calendars were organized and classes were formed (some schools regrouped frequently, others on a trimester basis, still others on a semester basis). As a result, ISR operations personnel (and some LEA personnel) had to work repeatedly to clarify the nature of the data being exchanged and to prepare rosters for verification by teachers and principals.

As part of this process, ISR researchers recruited 52 teachers and 17 principals to participate in a roster verification process using the web-based application described above. Each teacher and principal was provided a web URL and a password, and after accessing the URL through the assigned password, that teacher or principal could view and modify his or her assigned roster(s). A roster included: (a) a list of classes the teacher was listed as teaching in LEA-provided data, and (b) a list of students in each class. Teachers and principals were asked to check these rosters for accuracy and make any changes they found necessary to correct inaccuracies.

A total of 286 classes were listed on the rosters of the 52 participating teachers over the two (or three) time points when teachers were given roster data.[10] Teachers made relatively few changes to these class lists, but the changes that were made would affect the kinds of data provided to VAM vendors. For example, 2% of classes initially presented to teachers were recorded as being taught by another teacher; about 8% of classes were changed from one course code to another; and about 5% were changed to reflect team teaching arrangements. The most frequently occurring error in initial rosters was the grade level recorded for the class. In the roster pilot, more than 1 in 5 (or 22%) of classes were changed to reflect different grade levels. Once corrections were made to class lists, teachers examined the initial student rosters for each class. A total of 8583 students were listed as enrolled in the classes listed. Of these, only 2% (or 173) students were marked as not being in the class for which they were listed, and an additional 318 students who were not on that list in its original form were added to the course lists.

An interesting question is how long the roster verification process took teachers and principals to complete. To obtain an estimate of this, ISR researchers administered surveys to the teachers and principals who participated in the pilot. The survey data showed that the median teacher spent about 15-30 minutes on the rostering process each semester or trimester, with principals reporting about the same

[10] Teachers in semester systems were given rosters twice; teachers in trimester systems were given rosters three times.

time expenditure. Overall, teachers and principals generally found the web interface easy to use, reporting that the most difficult and time consuming task was adding students to rosters, and that the most difficult part of the rostering task to perform accurately was recording the dates at which students entered or left their classes.[11] Finally, the majority of teachers and principals reported that they would be willing to engage in this rostering process as part of the teacher evaluation process. In summary:

- ***The roster pilot showed that a roster verification process was possible in Michigan schools, that this process would uncover errors in the reporting of both teacher course assignments and student course enrollments, and that the verification process could be completed by most teachers and principals in 15-30 minutes time per semester.***

## Results of VAM Analyses

To this point, the chapter has described the various VAM analyses that can be conducted with MEAP data and the data processing issues involved in conducting such analyses. This section reports on some of the results of the VAM analyses.

The first results to note are these:

- ***When variance in adjusted achievement scores was decomposed into three components (students, classrooms, schools), the variance components in MEAP data differed from what is typically seen in large-sample student achievement data.***

- ***In addition, the teacher effects on adjusted student achievement estimated from value-added models using MEAP data tended to be at the lower end of what has been reported in other studies.***

Elsewhere, for example, analysts using covariate adjustment models like the ones estimated in the pilot have reported that the percentage of variance in adjusted achievement lying among classrooms varies from around 4-18% of the total variance in adjusted achievement.[12] These results can be compared to results from MEAP data by looking at the graphs presented on the next page. The data shown there come from a variance decomposition conducted by AIR in which variance in students' "adjusted" achievement was partitioned among students, teachers within schools, and schools. Because these data come from AIR's Model A, students' current achievement has been predicted from several years of students' prior achievement. The data show relatively small *classroom-to-classroom* variance in this adjusted achievement in Michigan schools. Indeed, in the MEAP data at hand, the variance in adjusted achievement among schools was always higher than the variance in adjusted achievement among classrooms within schools. This is unusual.

The largest classroom-to-classroom variance was in the 4th grade mathematics data, where about 63% of the variance in adjusted achievement was among students in classrooms, 11% was among classrooms within schools, and 26% was among schools. These data imply that two teachers who teach students with similar achievement histories but differ by a standard deviation in the distribution of teacher effects will produce a difference of about 3 MEAP scale score points in the average mathematics achievement score of their students. This translates to a δ type effect size of about .30 for adjusted MEAP scores and about .13 for unadjusted MEAP scores.[13] These "effect sizes" are not that unusual in VAM covariate adjustment analyses, except that in most analyses, there is more variance in adjusted achievement among classrooms than among schools.

In other grades, for both mathematics *and* reading, the data on the next page show that teacher effects were not as large as in 4th grade mathematics. Much more typical were the teacher effects in 6th grade mathematics, where variance among classrooms in adjusted achievement was 4% of total variance, two teachers a

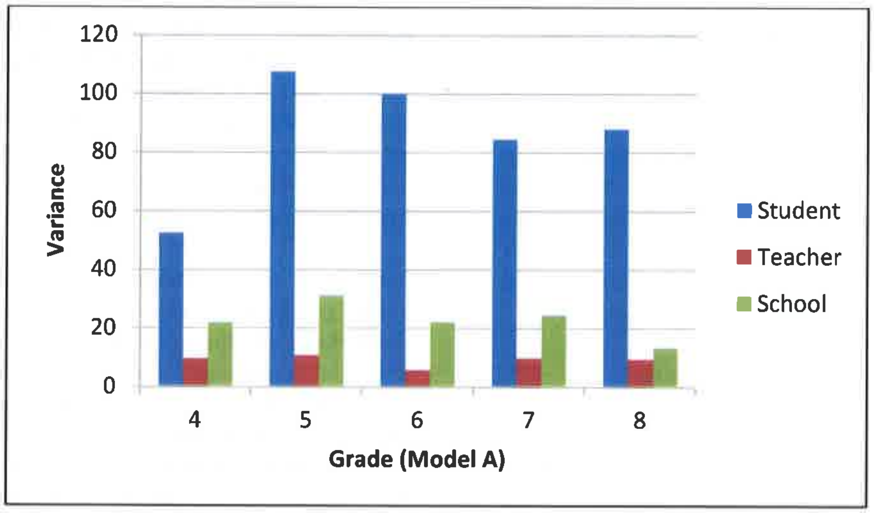
[11] This feature of the roster interface was added to assure that the roster interface would produce accurate data on the amount of time a teacher served as the instructor for any student.

[12] See, for example, Brian Rowan, Richard Correnti, and Robert Miller, "What Large-Scale Survey Research Tells Us About Teacher Effects on Student Achievement: Insights from the Prospects Study," *Teachers College Record*, 104(8), 2012.

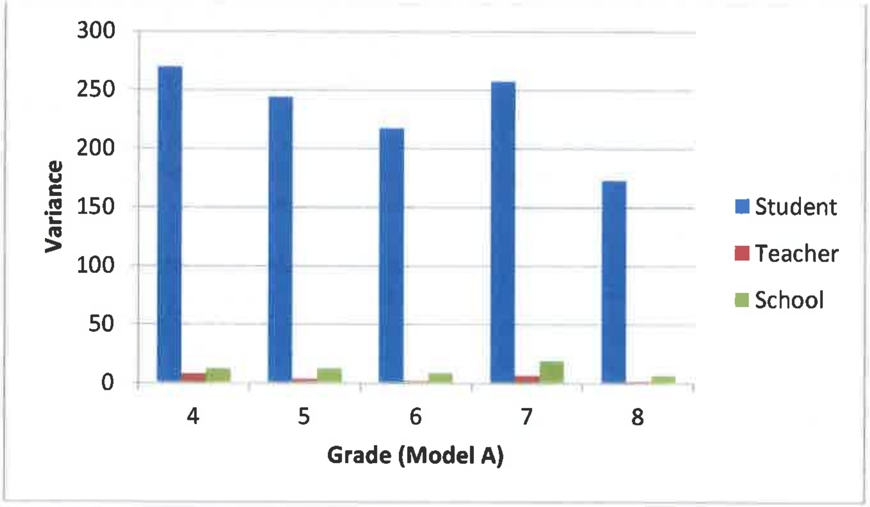
[13] We can compare this estimate to the SAS MRM estimate of teacher effects. The SAS MRM model, the reader will recall, is a layered model estimating teacher effects on students' *gains* in achievement (as measured by changes in Normal Curve Equivalent [NCE] scores above or below what would be predicted by prior achievement). Here, two 4th grade math teachers who differ by a standard deviation in the distribution of teacher effects (but who otherwise teach students with similar levels of prior achievement) are estimated to produce an average difference in student math gains of 3.88 NCEs.

## At a Glance: Percentage of Variance in Adjusted Achievement in MEAP Mathematics and Reading Lying Among Students, Teachers, and Schools

MEAP Mathematics Scores



MEAP Reading Scores



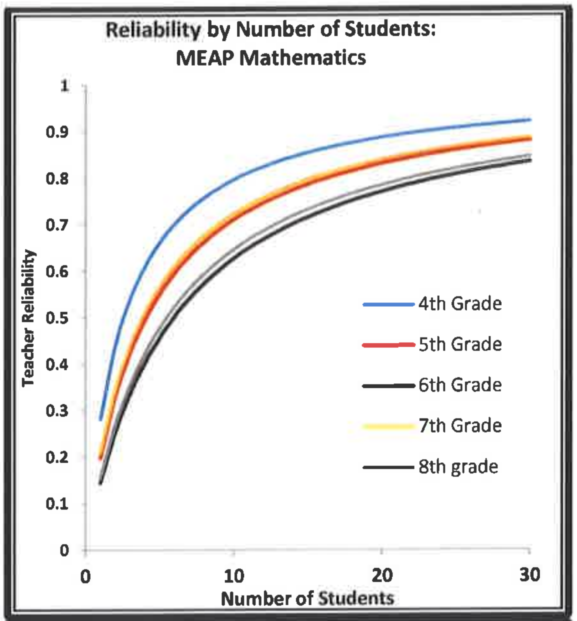
**AIR Model A is cross-classified hierarchical model with students, linked to multiple teachers, nested within schools. At level one, the model predicts a student's current achievement test score from several prior achievement test scores. Other levels of the model include random effects for teachers and schools.**

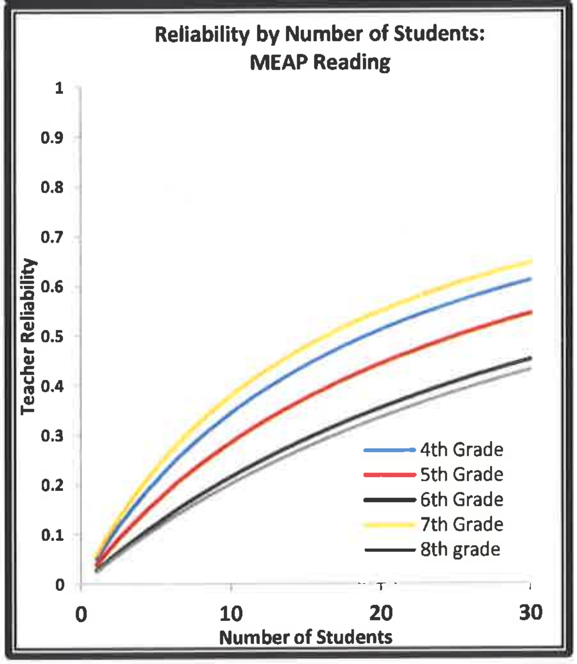
standard deviation apart in the distribution of teacher effects would produce a difference of about 2.4 MEAP scale points in mathematics achievement over the Fall-to-Fall period, and the $\delta$ type effect size would equal .21 for adjusted scores and .09 for unadjusted scores.

The teacher effects were similarly small in reading. For 4th grade reading, as an example, the percentage of variance in students' adjusted MEAP scores among classrooms was just 3%, two teachers a standard deviation apart in the distribution of teacher effects would produce a difference of a little less than 3 MEAP scale scores points in achievement over a Fall-to-Fall period, translating to a $\delta$ type effect size = .17 for adjusted MEAP reading scores and .10 for unadjusted MEAP reading scores.[14]

***The relative size of variance components in these analyses, coupled with the number of students linked to a teacher for a VAM analysis, affects the reliability of estimated teacher effects on student achievement.*** The graphs to the right show how this works for the MEAP data at hand. The graphs show that as the number of students linked to a teacher increases for a specific VAM analysis, the reliability of teachers' value-added scores increases.[15] The graphs also show that the absolute levels of reliability achieved by adding students depends also on the amount of variance in adjusted achievement that lies among teachers. Note, for example, that the reliability of teacher effect estimates for mathematics are always higher at grade 4 (where variance among teachers is largest) compared to reliability at other grades (where the teacher variance is smaller). Note also that teacher effect reliabilities are much higher for analyses in the area of mathematics versus reading. Again, this is due to the fact that there is more variance in adjusted mathematics achievement among teachers than there is variance in achievement among teachers in the area of reading.

## At a Glance: Reliability of Teacher Effect Estimate By Number of Students Linked to a Teacher





[14] For SAS, which is estimating teacher effects on students' *gains* in achievement (as measured by changes in Normal Curve Equivalent [NCE] scores above or below what would be predicted by prior achievement) the difference in NCE gains for teachers a standard deviation apart in the distribution of teacher effects is 3.1 NCE's; for 4th grade reading, the difference is 1.9 NCE's.

[15] The graphs are based on variance in student, teacher, and school variance components where "adjusted" student achievement is the dependent variable. Variance component estimates provided by AIR. The reliability coefficient presented here is discussed in Stephen W. Raudenbush and Anthony S. Bryk, *Hierarchical Linear Models: Applications and Data Analysis Methods*, 2nd Edition, Sage (2002: 230).

Overall, the average number of students with whom a teacher is linked in MEAP data is around 17. This suggest that the reliability of teacher effect estimates in the VAM analyses presented here are between .76 -.86 for analyses in mathematics and .30-.51 for analyses in mathematics.

Although these reliabilities appear reasonable (especially for mathematics), it should be noted that a lack of "relative" precision in VAM estimates can make it difficult to ascertain whether teachers' VAM scores differ from one another or from some established cut point that serves as a performance standard for evaluative purposes (a point discussed in more detail in Chapter 5 of this report). This problem occurs because the standard errors of VAM estimates are fairly large relative to the standard deviation in these estimates. An analysis presented by AIR of this issue suggests this problem is prevalent in *all* VAM estimates based on MEAP data, but is more exacerbated in estimates of reading teachers' value-added scores compared to esitmates of mathematics teachers' value-added scores.

A last set of findings from the VAM analyses concern the correlations among different VAM estimates of teachers' effects on students' achievement. This is an important issue for two reasons. First, this chapter has already noted that VAM vendors are prepared to estimate a variety of statistical models for clients, models that are based on different approaches and methodologies (e.g., growth versus covariate adjustment models). Moreover, a major controversy in the literature concerns whether or not to control for prior achievement only in these models, or to control for prior achievement plus other student characteristics, or to control for prior achievement, plus other student characteristics, plus "peer" effects.

The table to the immediate right shows the correlations among different value-added models. To produce the table, ISR researchers simply examined bivariate correlations of teachers' value-added scores to each other when estimated by the same vendor using different statistical models, and the bivariate correlations across different vendors. We only were able to do this for two vendors (SAS and AIR), but preliminary analyses with data from the third vendor (EA) suggest that results will not be different once the results of this vendor's analyses have been added to the table.

**At a Glance: Correlations of VAM Scores Within Vendors by Type of Model (for 5th Grade Math)**

| | SAS 1 | SAS 2 | SAS 3 |
|---|---|---|---|
| SAS 1 | | .94 | .95 |
| SAS 2 | | | .99 |
| SAS 3 | | | |

| | AIR 1 | AIR 2 | AIR 3 |
|---|---|---|---|
| AIR 1 | | .99 | .97 |
| AIR 2 | | | .98 |
| AIR 3 | | | |

SAS = layered model
AIR = covariate adjustment model, random teacher and school effects

1 = Controls only for students' prior achievement
2 = Controls for prior achievement + student characteristics
3 = Controls for prior achievement, student characteristics and peer effects

**At a Glance: Correlations of VAM Scores Across Vendor Models (for 5th Grade Math)**

| | AIR 1 | AIR 2 | AIR 3 |
|---|---|---|---|
| SAS 1 | .92 | .91 | .91 |
| SAS 2 | .88 | .90 | .92 |
| SAS 3 | .89 | .90 | .95 |

SAS = layered model.
AIR = covariate adjustment model, random teacher and school effects

1 = Controls only for students' prior achievement
2 = Controls for prior achievement + student characteristics
3 = Controls for prior achievement, student characteristics and peer effects

Once again, the table presents data for 5th grade mathematics, and again it is worth noting that the results shown in the table generalize to other grades. The main finding of note is that:

- ***The value-added estimates for a given teacher are highly correlated both within vendors as they control for more covariates and across vendors that use different statistical models to estimate teacher value-added scores.***

In fact, looking at the top of the table, one can see that within vendors, value-added scores for a teacher are highly correlated (greater than .94) as one moves from models that control only for prior achievement, to models that control for prior achievement plus other student characteristics, to models that control for these factors plus peer effects. Looking at the bottom of the table, one also can see that even though AIR and SAS use different statistical models to estimate teacher VAM scores, these estimates are almost always highly correlated (between .88 and .95).

It is important to note, however, that even when different VAM scores are as highly correlated across models as they are in the table on the previous page, some teachers' VAM scores *will change* from statistical model to statistical model.[16] Moreover, even slight changes in a teacher's VAM estimate can affect a teacher's annual evaluation, especially when cut points for assigning ratings are established near the center of the VAM score distribution. In evaluation systems that make these fine-grained, categorical distinctions among teachers, teachers with scores near the established cut points will be especially vulnerable to ratings changes that result from small changes in VAM scores produced by different statistical models.

Thus, although different statistical models produce highly correlated value-added estimates, value-added estimates *do* change across models in ways that can have important effects on teachers' annual evaluation ratings. For this reason, ISR recommends that if any education authority in Michigan plans to implement value-added modeling as part of its teacher evaluation process:

- ***A panel of educational and statistical experts should be convened to evaluate the technical quality of different approaches to value-added modeling. This panel can make recommendations about the value-added model(s) to be used to evaluate educators and about the vendor who will implement that approach in practice.***

[16] In an analysis conducted by EA, for example, about 20-40% of teachers' VAM scores changed by .20-.50 sd's (depending on subject and grade) as this vendor changed from models that include student covariates to models that included student covariates *plus* peer effects.

# Chapter 5: Setting Standards for Teacher Evaluation

This chapter addresses an issue discussed in Chapter 2 of this report. That chapter showed that there was no uniform standard for classifying teachers into the effectiveness ratings mandated by section 2(e) of PA 102 of 2011 and that, as a result, the percentage of teachers classified as "effective" and "highly effective" after annual evaluations varied widely from district to district, not because talent levels differed across districts, but because districts used different weighted formulae and set different cut points for assigning teachers to final effectiveness ratings.

In light of these findings, this chapter addresses two questions:

- How can districts go about setting performance standards for assigning effectiveness ratings to teachers?

- What percentage of teachers might end up being classified into different effectiveness ratings if a standards-based rating system is implemented?

The point of departure for addressing these questions is a discussion of two dimensions of the performance rating process: (a) measuring the "levels" of teacher performance; and (b) understanding the degree of statistical (un)certainty present in these measures. To illustrate how these two dimensions of performance rating inform the assignment of effectiveness ratings to teachers, this chapter first uses observation and VAM data to describe the levels of measured performance among Michigan teachers and then describes the degree of statistical (un)certainty associated with these measures. The chapter then describes two "standards-based" approaches to classifying teachers into effectiveness ratings. Both approaches take teachers' levels of performance into account in assigning effectiveness ratings. However, an initial method also takes into account the amount of statistical uncertainty surrounding a teacher's measured performance when assigning an effectiveness rating while a second approach does not. The chapter shows that the statistical uncertainty of teacher performance measures used in the pilot was high, and that because uncertainty was high, it was very difficult to assign teachers unambiguously into fine-grained ratings categories.

### Two Approaches to Performance Rating

Observation vendors in the pilot used "absolute" standards to judge teaching performance. In particular, each vendor's classroom observation tool had a set of items on which performance was to be rated, each tool had a rating scale that defined performance levels on these items, and each tool had an established scoring rubric to describe the behaviors and activities that justified assigning a performance level on an item to a teacher. FFT, for example, rated teachers on ten items (grouped into two domains called "the classroom environment" and "instruction"). In conducting a classroom observation, an observer assigned a rating on each item to teachers. In FFT (and also 5D) ratings were assigned at four performance levels: (1) unsatisfactory; (2) basic; (3) proficient; and (4) distinguished. TC had a similar rating system for assigning scores to items measuring its "four corners" of teaching effectiveness. It defined performance levels as: (1) novice; (2) developing; (3) proficient; and (4) expert. In rating classroom performance, then, observation vendors tended to rate the performance *level* of a teacher on an item.

VAM vendors took a different approach to performance standards. They assigned performance ratings to teachers based on two factors. The first was a teacher's measured impact on students' learning (i.e, a VAM score). The second was the degree of statistical (un)certainty associated with that estimate. When VAM vendors developed rating systems for teacher evaluations, they tended to ask the question: is a teacher's estimated VAM score significantly different from the statewide mean for teachers of similar students? To answer this question they used statistical procedures to create 95% (and 68%) confidence intervals around each teacher's estimated VAM score then looked at whether those confidence intervals included the statewide mean. If the confidence intervals overlapped with the statewide mean, teachers were generally labeled in VAM reports as having a VAM score that was "no different from average." However, when the confidence intervals did *not* overlap with mean, teachers were labeled as "significantly below average" or "significantly above average" depending on the absolute level of their score.

These examples show that: ***Performance ratings have two components. One component is an estimate of a teacher's level of performance (as measured by some measurement tool). The second is the degree of confidence one has in that estimate—as reflected in the confidence interval statisticians calculate for that estimate.***

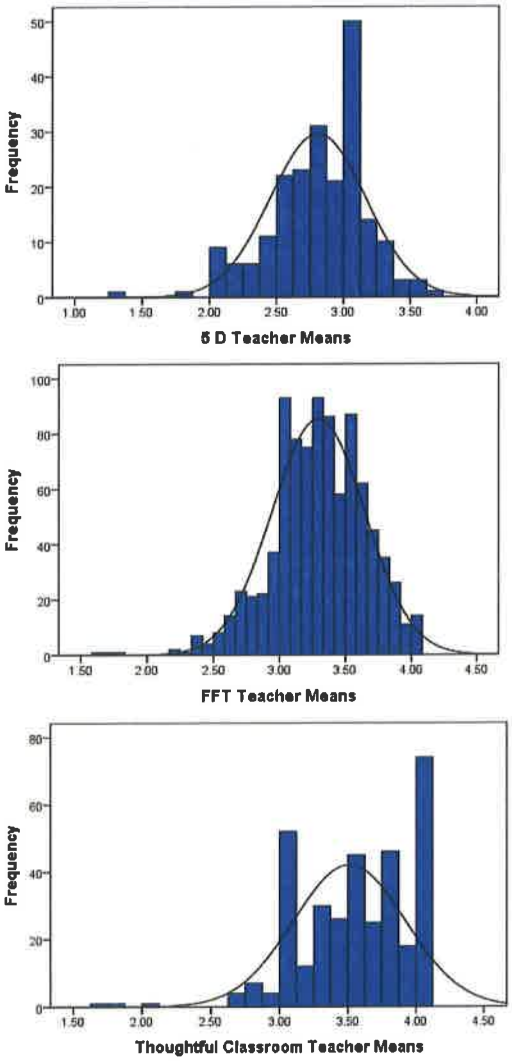
## Estimated Levels of Performance

An interesting descriptive question is what the "levels" of teaching performance were in the pilot study. For example, what did the observation data collected during the pilot tell us about how good the teaching was in pilot schools (at least as captured in the classroom observation tools)? And what did the VAM analyses tell us about how much student growth in achievement was produced by Michigan teachers (at least as captured by complex VAM estimates)?

One way to address these questions is to look at the distribution of teaching performance in pilot schools as measured by the classroom observation tools used in the pilot. In looking at these distributions, we are particularly interested in knowing the percentage of teachers who were rated as unsatisfactory, basic, proficient, and expert. These distributions are shown in the graphs to the right. Recall that each observation tool scored items on a scale of 1 to 4, where 1 was labeled as unsatisfactory by 5D and FFT and as novice by TC, where 2 was labeled as basic by 5D and FFT and as developing by TC, where 3 was labeled as proficient by 5D, FFT and TC, and where 4 was labeled as distinguished by 5D and FFT and as expert by TC. The scores assigned to teachers in the graphs to the right are simply average scores (across all items and occasions) for a teacher. No data are presented for the Marzano protocol because of extensive missing item data.

Looking at the graphs to the right shows that: ***The average score of most teachers on the observation tools tended to range from proficient (average score = 3) to distinguished (average score = 4),*** although this was not the case for scores on the 5D tool, where proportionally more scores were in a range from 2 to 3.5 (i.e., from basic to proficient). It is difficult to know if the difference between 5D scores and scores on the other tools is an effect of varying degrees of lenience (or severity) among districts using different tools, or if the 5D tool was perhaps measuring different dimensions of teaching than the other observation tools, or if the teachers in 5D districts were simply less proficient in teaching. ISR researchers are inclined to guess that the difference between 5D scores and the other score distributions reflects differences in the

### At a Glance: Distribution of Pilot Teachers' Scores on Pilot Tools



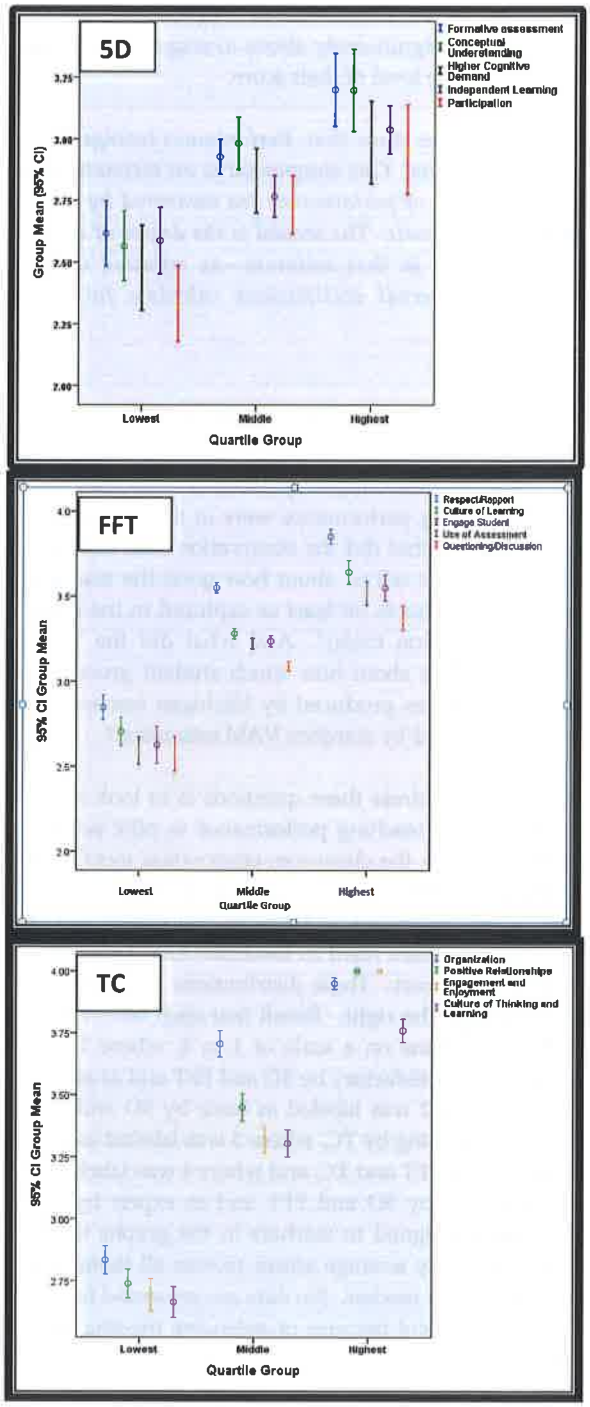
emphasis of the 5D protocol, which is heavily weighted to assessing cognitively demanding instruction.

***The average scores of teachers (shown on graphs on the previous page) obscure information about how teachers perform along specific dimensions of instructional practice.*** The graphs immediately to the right (on this page) show information at this level of detail. These graphs break teachers into three groups: teachers whose IRT scale scores on a tool were well-below the mean (i.e., teachers in the bottom quartile of the score distribution); teachers whose IRT scale scores were in the middle of the distribution (i.e., in the middle two quartiles), and teachers whose IRT scale scores were well-above the mean (i.e., in the top quartile of IRT scale scores).

What these graphs show is that: ***Teachers in the bottom quartile of measured performance—no matter what tool is being used—have item scores that are usually below "proficient" on the associated rating scales.*** One can also see that as we move up the performance distribution, item scores generally move up as well, so that teachers in the middle quartiles generally have scores that are centered a little above a proficient rating, and those in the top quartile of the performance distribution generally have scores approaching distinguished (or expert).

Importantly, the data to the right show another trend that is present within all performance groups: ***Teachers generally score higher on items that measure dimensions of the classroom environment (like "classroom organization" and "positive relationships with students") and score lower on items measuring important dimensions of instructional practice such as "developing a culture of thinking and learning" or "use of assessment techniques" or "questioning and discussion techniques."*** For example, the graph for item scores on the FFT tool is shown in the right middle. This graph shows item scores on the five best-fitting items on the FFT scale, but it still illustrates that, in all quartiles of the performance distribution, there is a fall-off in average item scores as one moves from the item measuring a classroom environment of respect and rapport to the item measuring a teacher's use of questioning and discussion techniques. A similar pattern can be found for the graph of TC items, where only "four corners" items are shown. On that

## At a Glance: Item Scores by IRT Score Quartiles



graph, there is a similar fall-off in item scores as one moves from the item measuring whether a teacher has a well-organized classroom to the item measuring how much the teacher maintains a culture of thinking and learning in his or her classroom. The only place where there is a departure from this trend is in the 5D graph. Here, the fall-off in item scores occurs from items measuring cognitive demand in instruction to items measuring formative assessment to items measuring participatory structures.

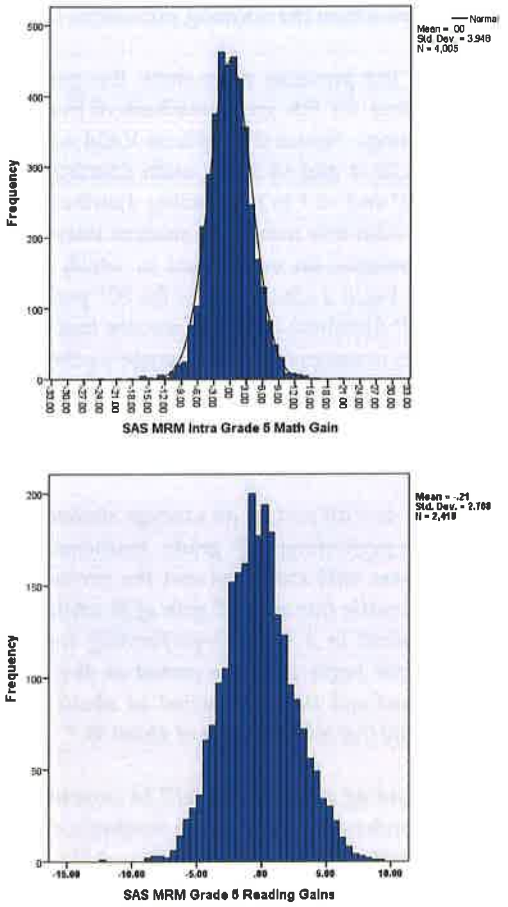
The graphs on the previous page provide interesting information. In particular, they suggest that even teachers at the very *top* of the performance distribution (as measured by the observation tools) have room for improvement, especially in critical areas of instruction, such as providing instruction of high cognitive demand, developing students' conceptual understanding, cultivating a culture of thinking and learning, and using good questioning and discussion techniques.

In summary, these graphs provide important information for setting performance standards in teacher evaluations. First, teachers in the lower quartile of the performance distribution have not achieved "proficiency" on many of the dimensions of classroom teaching rated by the observation tools. Second, even teachers at the top of the ratings distribution have room to improve, especially in key areas of classroom practice.

We also can look at VAM score distributions to get a picture of the levels of performance of teachers in promoting student learning. As we have seen, VAM scores are inherently "relative" measures of performance. But: ***VAM analyses provide information on how much teachers who teach similar students differ from one another in their "absolute" impact on student learning gains.*** The graphs to the right, for example, provide information about the magnitude of teacher impacts on student achievement gains over the Fall-to-Fall MEAP testing interval.

To interpret this graph, the reader should know several things. First, the data are from a VAM analysis conducted by SAS using the MRM model that controls only for prior achievement. Second, SAS measures student gains in achievement as changes in a student's normal curve equivalent score relative to changes made by the rest of the statewide sample. Using this relative metric, a positive score indicates

## At a Glance: Relative Impact of Teachers on Grade 5 Math Gains



**These are histograms of the distribution of estimated teacher effects on students' gains in 5th grade mathematics and reading (from the SAS VAM analysis using the MRM model 1). The average teacher impact is centered at zero. One can see from these graphs that students of a teacher one standard deviation above the mean in the performance distribution for 5th grade mathematics teachers will end the year about 4 NCE's above students in the average teacher's class. Students of a teacher one standard deviation above the mean in the performance distribution for 5th grade reading teachers will end the year about 3 NCE's above students in the average teacher's class.**

"above average" gains (a student has made larger gains than the norming population), and a negative score indicates "below average" gains (a student has made smaller gains than the norming population).

The graphs on the previous page show the performance distribution for 5th grade teachers of mathematics and reading. Notice that 68% of VAM scores are between about -4 and +4 in the math distribution and between -2.7 and +2.7 in the reading distribution. To understand what this means for student learning, it is useful to imagine an experiment in which two similar students begin a school year at the 50th percentile of the MEAP distribution. Now, assume that one of these students is assigned to the average teacher in Michigan's VAM score distribution while the other is assigned to a teacher who is one standard deviation above the mean in that distribution. Then:

- ***Over a Fall-to-Fall period, an average student in an average-performing 5th grade mathematics teacher's class will start and end the period at the 50th percentile (for an NCE gain of 0) while an average student in a superior-performing teacher's class will begin the time period at the 50th percentile and end the time period at about the 57th percentile (for an NCE gain of about 4).***[17]

***Differences in reading achievement will be somewhat smaller.*** The student in the average teacher's class again ends the time period at the 50th percentile (for an NCE gain of 0), whereas the student in the superior teacher's class ends the time period at about the 55th percentile.[18]

Again, this information is useful for setting performance standards in teacher evaluation. We already know that the average teacher in Michigan has a VAM score of 0 (indicating that his or her students are experiencing academic growth on test scores at the same pace as the average student in Michigan). So one question central to setting performance standards for VAM scores would be: How far behind do a teacher's students need to fall before we decide that the teacher is, for example, ineffective? If a teacher's VAM score is -4 in mathematics, the average student would drop from the 50th percentile on the MEAP mathematics test to the 42nd percentile over a Fall-to-Fall period; if a teacher's VAM score is -6, the average student would drop from the 50th percentile to about the 39th percentile; if a teacher's VAM score was -8, the average student would drop from the 50th percentile to about the 36th percentile. Evaluators need to decide which of these performance levels should be used to signal "ineffective" teaching.[19]

## Imprecision in Teacher Performance Estimates

Looking at a teacher's score from an observation instrument or VAM analysis is an important part of the evaluation process. But an important concept from measurement theory is this: ***Any measure of a teacher's performance, whether from a classroom observation or a VAM model, is an estimate of that teacher's performance, and estimates come with uncertainty.*** This uncertainty arises from errors of measurement, of which there are many in the measurement of teaching performance. For example, we have already seen in previous chapters that errors in measurement from classroom observations arise because teacher's estimated performance can vary from occasion to occasion, from observer to observer, and from item to item. With VAM scores, the primary source of error variance is the number of students whose achievement is being considered in the VAM estimate.[20]

The usual way uncertainty in measurement is quantified by measurement experts is through the standard error of measurement (SEM). One way to understand the SEM is to see it as an estimate of how much repeated measures of a person on the same instrument will be distributed around that person's "true" score. In general, SEM's are larger when reliability is lower, simply because measurement errors (not changes in true performance) are producing score variance. As it turns out, the SEM is related not only to measurement reliability, but also to the "confidence intervals" that statisticians set around estimates. A confidence interval can be thought of informally as an estimated

[17] We can look at the reverse case, where one student gets the average teacher and another gets a teacher a standard deviation below the mean of the teacher performance distribution. Then the student of the average teacher ends the year at the 50th percentile while the student assigned to the low performing teacher falls to about the 42nd percentile.

[18] Researchers often look at more extreme differences in the teacher performance distribution. For example, if one of our 50th percentile students was assigned to a teacher two standard deviations above the mean of the teacher performance distribution, that student would end the year at the 64th percentile.

[19] The same problem arises in deciding who to classify as "highly effective." How far (above average gains) must a teacher's students be boosted to see if he or she is to be classified as highly effective?

[20] The precision of VAMs can also be affected by measurement errors in the independent variables—especially prior test scores.

range of values which is likely to include the unknown "true" score of a person (given a sample of data). A 95% confidence interval is bounded by 1.96 SEMs on both sides of the estimate, and a 68% confidence interval is bounded by 1 SEM on both sides of the estimate.

***The SEM and confidence intervals for performance measures are important to the evaluation of teachers for several reasons.*** Suppose, for example, that an education authority wants to know how confident it can be that a teacher has met a standard of classroom teaching performance required for tenure. For example, suppose the education authority has said a teacher must be "proficient" in teaching (i.e., have a score of 3 on the FFT framework) in order to obtain tenure. Now suppose that a teacher has been scored at 2.5 on a series of classroom observations (i.e., the teacher's score is "basic" on the scoring rubric, but not yet "proficient"). The question an education authority might want to address is how confident it can be that this teacher is truly *not* proficient (i.e., does not have a score of 3). One way to address this question is to put a 95% confidence interval around the estimate of 2.5 and see if it includes a score of 3 (the standard required for tenure). Of course, there is nothing sacred about a 95% confidence interval. An education authority might, for example, only want to be 68% confident in its decision, in which case it would use a 68% confidence interval.[21]

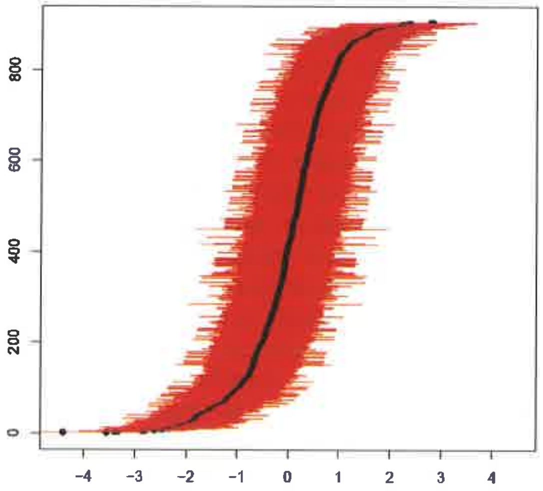
The use of such confidence intervals in performance measurement is quite common. Confidence intervals are widely used in student assessments to determine whether students have (or have not) reached particular performance levels (such as proficient). VAM vendors also use confidence intervals when they classify teachers into performance categories. In particular, VAM vendors often use 95% (and 68%) confidence intervals on a teacher's VAM score to see if a teacher's estimated VAM score overlaps with the sample mean. When a teacher's estimated VAM score is below the mean and the 95% confidence interval for the estimate does not include the mean, VAM vendors say that teacher is "significantly below average" in performance. Or, if the teacher's estimated VAM score is above the mean and a 95% confidence interval does not overlap with the mean, they might call that teacher "significantly above average."

Note, however, that VAM vendors are using the *mean* of the score distribution as the reference standard, and this brings our discussion back to the critical problem of how to choose a particular "level" of performance as the reference standard for personnel classification. An education authority, for example, might want to screen out teachers whose VAM scores are 1.5 standard deviations *below* the mean (which would mean the average student in these teachers' classes would experience decrements of about 6 NCE's over a year compared to similar peers). It is only *after* a standard has been set that confidence intervals come in handy. Confidence intervals tell evaluators the "chances" that an employee's measured performance overlaps with, or is above or below, a particular standard of performance that has been set in advance.

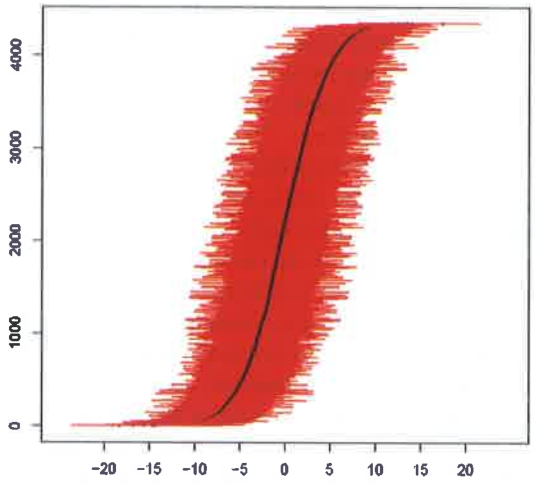
***An important point about VAM and teacher observation scores is that they have very large standard errors of measurement relative to the distribution of estimated performance scores.*** The graphs at the top left of the next page, for example, show the 95% confidence intervals for scores on one of the observation tools that MCEE piloted (FFT) *and* for one grade/subject combination of VAM score estimates (4th grade math using SAS MRM estimates). The X (or bottom) axis of each graph shows the range of (observation or VAM) scores. The Y (or vertical axis) shows the cumulative number of teachers. The black dots on the graph are scores for individual teachers. The red lines running through each black dot are the 95% confidence intervals for each score. Note that the 95% confidence interval is different for each score. For observation data, that occurs because different teachers have been observed on different numbers of occasions, with more or less observer error, and during that time, perhaps scored on more or fewer items. For VAM data, that is because teachers have been linked to different numbers of students using TSDL data, and teachers whose VAM scores are estimated from fewer students will have larger confidence intervals.

[21] The same problem can be framed as one of deciding about giving rewards to teachers. For example, suppose an education authority sets a standard of 3.5 on FFT for award of "master teacher" status. Then the problem is once again to place a confidence interval around a candidate's estimated score to see if it is "significantly" above the required standard.

### At A Glance: 95% Confidence Intervals for IRT Scale Scores on the FFT Observation Tool



### At a Glance: 95% Confidence Intervals for SAS MRM VAM Scores for 5th Grade Mathematics Teachers



Now, look at the top graph to the left. This graph shows the estimated observation scores of teachers who were observed with FFT (in IRT scale score points) and the 95% confidence intervals for these estimates. ***What can be seen from the graph is that the 95% confidence intervals for teacher observation scores are large relative to the distribution of IRT scale scores.*** As an example, notice that the 95% confidence interval for a teacher whose FFT score is 2 on the graph runs from about +3 to about 0. This makes it very difficult to confidently distinguish teachers' observation scores from one another and (as we are about to see) very difficult to confidently ascertain whether a given teacher falls above or below some cutoff for meeting a particular standard of performance.

Next, look at the graph at the bottom left of this page. This graph shows the estimated VAM scores for 4th grade mathematics teachers (using the SAS MRM model) as well as the 95% confidence intervals for these estimates. Bearing in mind that with SAS VAM scores, the mean of the performance distribution is 0, and that the scale indicates the relative boost or decrement to NCE scores that the average student in a class would experience over a Fall-to-Fall period, we can look at how precisely any teacher's VAM score is estimated. ***What can be seen from the graph at the bottom left is that the 95% confidence intervals for teacher VAM scores are large relative to the distribution of scores.*** As an example, the 95% confidence interval for a teacher with a VAM score of +5 runs from about +10 to -5. Again, this makes it difficult to confidently distinguish teachers' VAM scores from one another and (as we shall see later) to confidently ascertain whether a given teacher falls above or below some cutoff for meeting a particular standard of performance.

### Taking Imprecision into Account in Making "High Stakes" Personnel Decisions

To this point, the chapter has demonstrated two elements of a standards setting process for teacher evaluation. As a first step, we looked at performance distributions and deliberated about the absolute score levels that would determine assignment of a rating category (like "ineffective") to a teacher. We have also looked at how certain decision makers can be about a teacher's measured level of performance (by examining standard errors of measurement [SEMs]). These SEMs, it will be recalled, were quite large relative to the distribution of scores.

***The lack of relative precision of teacher performance measures has important implications for the classification of teachers into the four effectiveness groups defined by section 2(e) of PA 102 of 2011.*** It should go without saying that the task of assigning teachers into ratings categories comes with real stakes for teachers—especially if a teacher is to be assigned to the "ineffective" category. By law in Michigan, if a teacher is classified as ineffective three years in a row, that teacher must be dismissed. Education authorities should therefore exercise care in making this classification. The law also requires that public education agencies classify teachers into three additional categories (minimally effective, effective, and highly effective). But:

- ***The challenge in classification of teachers into final effectiveness ratings is that confidence intervals around observation and VAM score estimates are large relative to the performance distribution. This makes it difficult to make classification decisions with a high degree of statistical confidence.***

To see this, we now explore an approach to assigning teachers to ratings categories using confidence intervals to assess the degree of confidence that decision makers can have about whether a given teacher's job performance does or does not meet some consequential performance standard.

In the example, we assume that an education authority is deciding whether a teacher exceeds the performance level needed to be classified as "ineffective." As we have seen, to make this decision, the education authority needs to set some absolute standard required for this decision. In the following examples, we arbitrarily assume that the education authority has set a score of 2 (or "basic") on the FFT scale as the cut point that must be exceeded to avoid classification as an ineffective teacher (which translates to a score of about 1.5 standard deviations below the sample mean). In addition, we shall assume (again arbitrarily) that the education authority has said that any teacher must have a VAM score above -6 to exceed the threshold for being classified as ineffective on the basis of VAM scores (this again translates to a score that is about -1.5 sd's below the mean of the VAM score distribution). In the scenario, then, the education authority under discussion has already set its performance standards in advance.

Now, let us suppose that in making decisions, the education authority wants to be *confident* in its decisions about teachers. In particular, suppose it wants to be fairly certain that when it declares a teacher whose score is below the relevant cut point for being classified as "ineffective" that there is a strong chance this classification reflects a teacher's true score—not measurement error. To gain perspective on this issue, the education authority can set a confidence interval around the teacher's estimated score and examine whether that confidence interval overlaps with the standard being used to make the decision. If the confidence interval overlaps with that standard, the education authority cannot be confident that the teacher's true score is really below the cut point, but if the confidence interval does not overlap with the cut point, it will have more confidence in its decision.

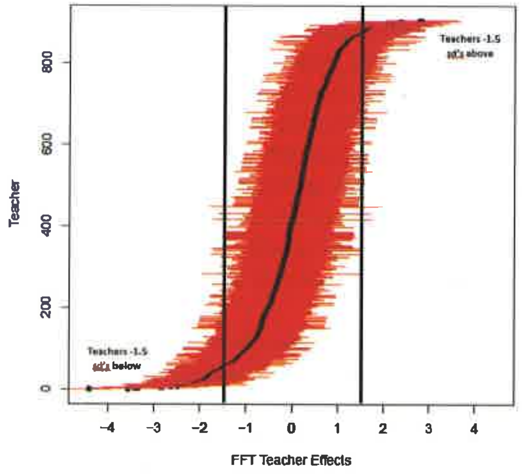
The figures on the next page show how this kind of decision making works for FFT scores (top graph) and VAM scores (bottom graph). These are the same graphs shown previously, except now we have drawn a vertical line running through the left hand side of the X (or bottom) axis of the graph to show the cut point (or standard) above which teachers must score to avoid being classified as "ineffective." All scores to the left of this line are below the established cut point for being classified as "ineffective" on the basis of an FFT or VAM score, and all scores to the right of that line meet or exceed the standard.[22]

Now, look once again at the red lines running through the scores in the region of the cut point on the graphs. These are the 95% confidence intervals of estimated FFT and VAM scores. What is immediately evident from the graphs is that not *all* teachers with FFT or VAM scores below their respective cut points can be said with 95% confidence to be "ineffective" since many 95% confidence intervals for scores to the left of the applicable cut point run through that cut point. In fact, using 95% confidence intervals, only teachers at the very extremes of the distributions (with scores below -2.5 on FFT and scores below -10 on the VAM measure) can be labeled as "ineffective" with 95% confidence.[23] Moreover, we also can see from the graphs that many scores above the cutoff

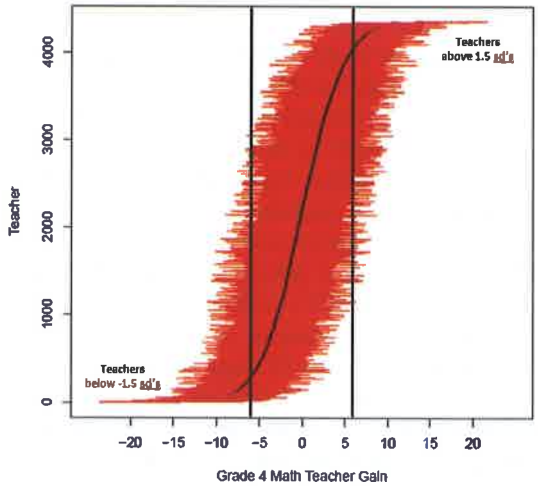
[22] The line on the right hand side of the graph is a cut point for determining whether a teacher can be classified as highly effective.

[23] We can take a more optimistic scenario and suppose the education authority wants to identify "highly effective" teachers. The situation here is the same. Only a handful of teachers whose FFT

### At a Glance: Using Scores and Confidence Intervals to Classify "Ineffective" Teachers Using FFT Scores



### At a Glance: Using Scores and Confidence Intervals to Classify "Ineffective" Teachers Using VAM Scores



line (i.e., to the right of the vertical line) also have 95% confidence intervals that intersect with the cut point. Thus, decision errors can be made on *both sides* of the cut point.

### The Problem of Joint Classification

To this point, we have discussed classification problems using a single performance measure at a time. But, PA 102 of 2011 calls for the use of *multiple* performance measures to classify teachers into section 2(e) ratings categories. For that reason, this section turns to the problem of *joint* classification, that is, classification that involves the use of more than one performance measure to make personnel decisions.

There are a variety of ways to make classifications using more than one performance measure. Perhaps the most common approach is to form a linear composite of the two indicators. This is in fact what most districts in the MCEE pilot did, and it is the method implicit in PA 102 of 2011, which calls for more weighting to be given to student learning measures in assigning effectiveness ratings to teachers. [24] We do not discuss this approach now (although we will discuss it in a technical report). Rather, this chapter focuses on a method of joint classification that is consistent with MCEE's recommended approach to assigning effectiveness ratings to teachers (see page 23 of *Building an Improvement-Focused System of Educator Evaluation in Michigan: Final Recommendations*, July 2013).[25] Unlike the MCEE approach, however, the approach to joint classification illustrated in this report takes both estimated scores *and* confidence intervals into account in making classification decisions.

MCEE's final report called for education authorities to use two primary measures to assign final effective-

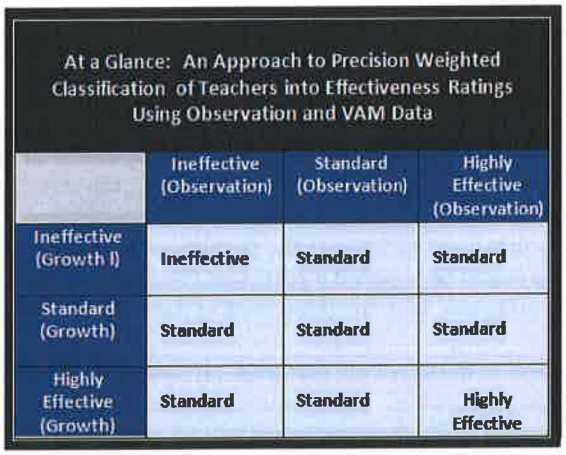
scores are above +1.5 sd's can be said to be above the established standard with 95% confidence.

[24] As an example, a simple way to combine the two measures into a linear composite would be to create a simple formula like: Composite score = $w_1$(Observation Score) + $w_2$(VAM score), where the w's are weights decision makers want to attach to different scores and observation and VAM scores are on the same scale.

[25] MCEE's approach has often been called a "conjunctive" approach to decision making whereas approaches using linear composites are called "compensatory" approaches. In a forthcoming technical report, ISR researchers will show how the composite weighting approach differs from the MCEE approach in failing to articulate clear performance standards. The forthcoming report also will show how to deal with measurement errors when developing linear composites of performance measures.

ness ratings to teachers: (1) teacher observation scores; and (2) student growth scores. The table immediately to right shows how MCEE's approach might work (although it is not *exactly* the same as the table shown in MCEE's final report). In the table to the right, ISR researchers are assuming that Michigan has transitioned into a fully developed system of teacher evaluation in which two kinds of data are readily available to decision makers. The first is a score from a state-approved classroom observation tool. In the empirical illustrations presented below, ISR researchers will assume that each teacher has been observed on about 4 occasions per year and that decision makers have an IRT scale score for each teacher on this instrument, along with a standard error for that score. The second score comes from a fully-developed system that provides VAM scores. In the empirical illustrations to follow, ISR researchers assume that decision makers are using VAM scores generated from the SAS MRM model for teachers of 4th grade mathematics, along with the SEMs for each score.

To construct the table to the right, ISR researchers assigned teachers to cells in the 9-fold table taking 95% confidence intervals into account. In this approach, they first assigned ratings by grouping teacher separately on observation and VAM scores into those that were above or below the respective cut scores for being classified as "ineffective" or "highly effective" on each measure (making sure that a teacher was classified as ineffective or effective only if the 95% confidence region for scores of teachers did *not* overlap with the relevant cut score).[26] After doing this, ISR researchers produced three groups of teachers on each ratings category—a group of ineffective teachers on a rating dimension (like FFT), a group of highly effective teachers on that rating dimension, and a remaining group of teachers who could not be classified into either of these ratings and received a rating of "standard." ISR researchers then cross-tabulated these separate ratings to produce the 9-fold table shown above.[27]



At a Glance: An Approach to Precision Weighted Classification of Teachers into Effectiveness Ratings Using Observation and VAM Data

| | Ineffective (Observation) | Standard (Observation) | Highly Effective (Observation) |
|---|---|---|---|
| Ineffective (Growth I) | **Ineffective** | **Standard** | **Standard** |
| Standard (Growth) | **Standard** | **Standard** | **Standard** |
| Highly Effective (Growth) | **Standard** | **Standard** | **Highly Effective** |

An important practical question concerns the number of teachers who can be expected to fall into the various cells of this 9-fold table (given what we know about the distribution of performance scores from the pilot research). Put differently, under the decision rules just described, how many teachers will be classified as "ineffective," how many will be classified as "highly effective," and how many will fall into the "standard" category? The answer to these questions has real cost implications. For example, classifying teachers as ineffective increases supervision costs (as a matter of law and good employment policy), and when such classifications lead to dismissal, districts face the costs associated with recruiting new teachers and training them. Classifying teachers as highly effective also affects supervision costs, since under PA 102 of 2011 several aspects of the annual evaluation process (e.g., annual classroom observations) are eased for highly effective teachers (thus reducing principals' evaluation workloads). A large group of teachers in the "standard" classification means more teacher observations, more teacher conferences, and more reporting for principals, since teachers in this group must be evaluated annually.

ISR researchers conducted a simulation study to estimate how many teachers would end up in the various cells of the table above, and to see how those numbers would change as more years of data were used on each teacher. The goal of using multiple years of data

[26] Note the shift in language from confidence interval to confidence region. A confidence region is simply a multivariate extension of the confidence interval.

[27] It is worth noting that other approaches to classification could be taken. For example, we could generate the four category rating system described in section 2(e) of PA 102 if we: (1) label any teacher who was ineffective on both scores as "ineffective"; (2) label any teacher who was ineffective on any one score dimension as "minimally effective"; (3) label any teacher who was neither ineffective or effective on either criteria as "effective"; and (4) label any teacher who was effective on both scores as "highly effective." This would produce the four ratings categories listed in section 2(e) of PA 102 and still be based on rigorous statistical procedures.

was to increase measurement precision. In constructing this simulation, ISR researchers used data on the distributions and standard errors of VAM and FFT scores from pilot data and then assumed a correlation of $r$ = .40 among the two performance measures to create a simulated data set of 905 teachers like those who would be found in Michigan.[28] The simulation was then used to forecast the percentage of teachers who would be assigned a particular effectiveness rating using one, two, and three years of data.

A table at the top of the next page presents the results of this simulation. Here, teachers whose VAM *and* FFT scores fell below the cutoff (of -1.5 sd's) on both measures with 95% confidence were labeled as ineffective and placed in the red-shaded cell of the table. In addition, teachers whose VAM *and* observation scores were above the cutoff (of +1/5 sd's) on both measures with 95% confidence were labeled as highly effective and placed in the green cell of the table. The remainder of teachers were placed in the buff-colored cells and labeled as "standard."

***The simulated data (on the next page) suggest that using 3 years of data—with the average teacher having about 12 FFT observations and a VAM based on 51 students—.5% of teachers will be classified as ineffective using the 95% confidence region ISR set for decision making and no teachers will be classified as highly effective using a 95% confidence region.*** The reason for these very low percentages at these extremes of the joint distribution is the imprecision of both the classroom observation measures and VAM measures relative to their respective distributions.

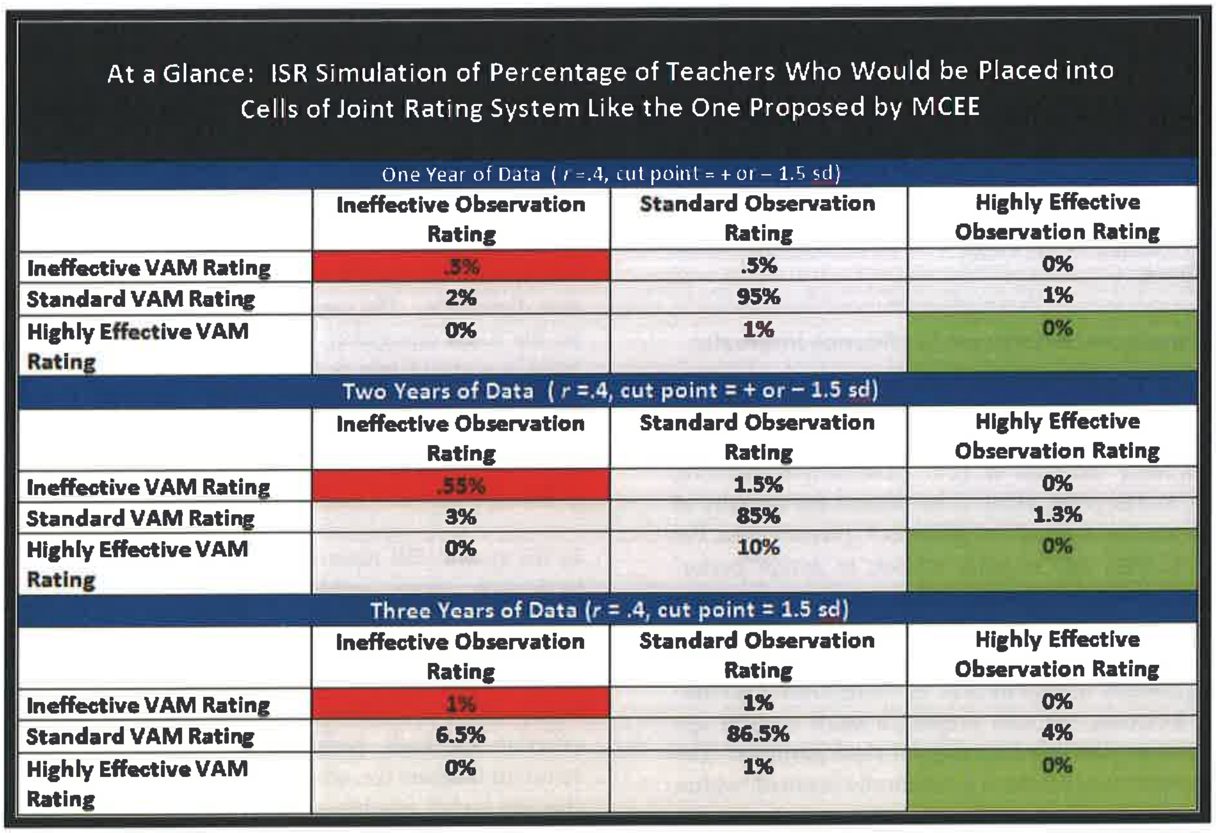
There are several ways to change the percentages of teachers in the cells of the table just discussed. For example, keeping a 95% confidence region:

- An education authority might want to give more weight to VAM scores in classification (as implied by PA 102 of 2011). To do this, it could assign a rating of ineffective overall to any teacher who was classified as ineffective on the VAM measure (with 95% confidence) *no matter what* the teacher's observation rating. The simulation showed that this would increase the percentage of teachers classified as ineffective from .5% to 1% of all teachers.
- Alternatively, an education authority could set a lower threshold for being classified as "ineffective" or "highly effective." For example, the standard could be set at -1.0 sd's for being classified as ineffective on an observation score or a VAM score, and +1.0 sd's to be classified as "highly effective." Using this decision rule, ISR's simulation found that 2% of teachers could be classified as ineffective and 1% of teachers as highly effective with three years of data.
- Alternatively, an education authority could use the cut point often used in VAM analyses and use the mean of the distribution as a standard for classifying teachers as "ineffective" or "highly effective."[29] Using this approach, any teacher whose VAM and FFT scores and 95% confidence region for those scores was below the *mean* would be classified as ineffective, and any teacher whose VAM and observation scores were significantly above the mean would be classified as highly effective. Using these decision rules, about 23% of teachers would be classified as ineffective and about 28% of teachers as highly effective, leaving about 44% of teachers in the "standard" classification. This approach, however, is clearly unsustainable, for few districts could afford the replacement and supervision costs of this approach.
- Finally, an education authority could change the confidence region used in decision making. For example, ISR researchers explored the possibility of setting a 68% confidence region instead of a 95% region. The results using 3 years of data are shown in the bottom table on the previous page. That approach labels about 1.2% of teachers as ineffective and .3% as highly effective.

The main point of the discussion is this: ***Using a reasonable set of performance standards, very few teachers in Michigan can be rated with 95% (or 68%) confidence as being ineffective or highly ineffective. Instead, most teachers can only be rated with 95% (or 68%) confidence as "standard" teachers (who are neither ineffective nor highly effective).***

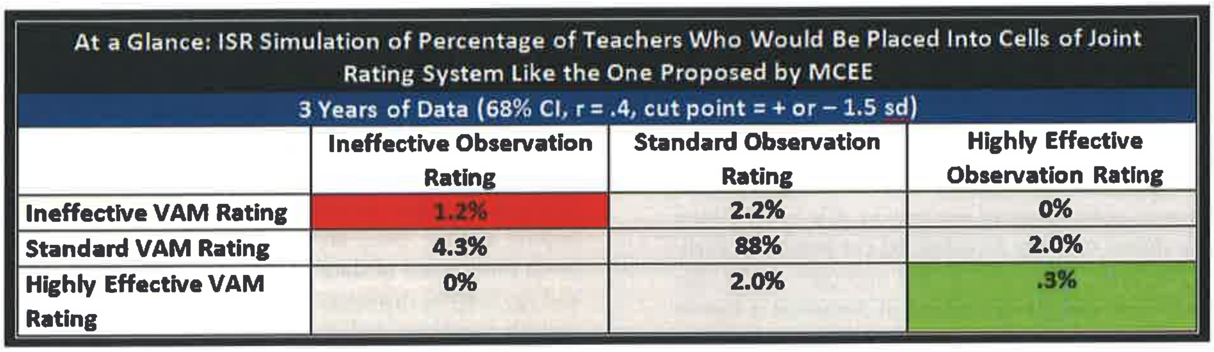
[28] The assumption that VAMs and observation scores are correlated at $r$ =.40 comes from data previously analyzed by ISR researchers (see Rowan and Raudenbush, op cit.). However, the correlation could, in fact, be higher or lower in Michigan.

[29] To be clear, VAM vendors do not recommend using this approach as a standard for *consequential* personnel decisions. They simply use accurate labels like "above average" and "below average" as diagnostic indicators and they are clear to signal the statistical meaning of their classification system.



At a Glance: ISR Simulation of Percentage of Teachers Who Would be Placed into Cells of Joint Rating System Like the One Proposed by MCEE

| One Year of Data ( *r* =.4, cut point = + or – 1.5 sd) | Ineffective Observation Rating | Standard Observation Rating | Highly Effective Observation Rating |
|---|---|---|---|
| Ineffective VAM Rating | .5% | .5% | 0% |
| Standard VAM Rating | 2% | 95% | 1% |
| Highly Effective VAM Rating | 0% | 1% | 0% |
| **Two Years of Data ( *r* =.4, cut point = + or – 1.5 sd)** | **Ineffective Observation Rating** | **Standard Observation Rating** | **Highly Effective Observation Rating** |
| Ineffective VAM Rating | .55% | 1.5% | 0% |
| Standard VAM Rating | 3% | 85% | 1.3% |
| Highly Effective VAM Rating | 0% | 10% | 0% |
| **Three Years of Data (*r* = .4, cut point = 1.5 sd)** | **Ineffective Observation Rating** | **Standard Observation Rating** | **Highly Effective Observation Rating** |
| Ineffective VAM Rating | 1% | 1% | 0% |
| Standard VAM Rating | 6.5% | 86.5% | 4% |
| Highly Effective VAM Rating | 0% | 1% | 0% |



At a Glance: ISR Simulation of Percentage of Teachers Who Would Be Placed Into Cells of Joint Rating System Like the One Proposed by MCEE

3 Years of Data (68% CI, r = .4, cut point = + or – 1.5 sd)

| | Ineffective Observation Rating | Standard Observation Rating | Highly Effective Observation Rating |
|---|---|---|---|
| Ineffective VAM Rating | 1.2% | 2.2% | 0% |
| Standard VAM Rating | 4.3% | 88% | 2.0% |
| Highly Effective VAM Rating | 0% | 2.0% | .3% |

*From this perspective, it also should be obvious that the major emphasis of PA 102 of 2011 cannot be on classifying teachers into a set of fine-grained performance ratings, for the tools used in the pilot simply do not have the needed precision. Instead, as MCEE pointed out in its final report, the main goal of conducting teacher evaluations under PA 102 has to be to produce improved teaching and learning.* The main way the law enables such learning is through the provision of feedback to teachers about teaching quality. We have already seen from data presented in Chapter 2, for example, that many teachers and most principals think the observation measures they piloted provided accurate performance information (although both groups were somewhat skeptical about the worth of information provided from standardized tests). Moreover, both principals and teachers reported that conferencing gave them an opportunity to come together and discuss performance measures and improvement steps in a satisfying way. Thus, to say that pilot tools cannot make fine-grained distinctions among teachers is not to say they are useless. ***For the vast majority of employees, the main import of an evaluation system will be to stimulate employee improvement, not to make a consequential personnel decision.***

## Classification Without Confidence Intervals: Simple Ranking Systems

An approach to personnel classification that relies on confidence intervals is both technically-demanding and, at this point, likely to be beyond the capacity of all but a few local school districts.[30] Nevertheless, PA 102 of 2011 still requires schools to assign performance ratings to teachers and use them to: (a) determine teacher dismissals; and (b) make reductions in force. ***In light of PA 102's requirement that annual effectiveness ratings be used in consequential personnel decisions, we now propose a much simpler approach to joint classification for these purposes.*** The approach we describe is scientifically justified, within the capacity of all public education agencies to implement, and meets the requirements of section 2(e) of PA 102 of 2011.

The approach involves ranking teachers on their combined observation and VAM scores in a very coarse way, where by "coarse," we mean "made up of large pieces." To illustrate how this coarse ranking system works, suppose an education authority once again has assembled VAM and observation scores and once again assigned teachers to cells in the 3-by-3 table shown on page 41 using the cut points set earlier. The main difference in the approach to be discussed now and the approach just discussed is that in approach we describe next, measured scores on VAMs and observations are used for decisions, but confidence intervals for scores are not calculated or used. Instead, teachers are simply assigned to ratings categories based on measured scores.

The coarse ranking system ISR researchers envision would begin by assigning teachers a score of 1-3 on each performance dimension separately. For example, if a teacher is classified as "ineffective" on the classroom observation metric, that teacher would get a score of one, if the teacher was assigned a rating of "standard" on this dimension, the teacher would get a score of two, and if the teacher was assigned a rating of "highly effective" on the classroom observation component, the teacher would get a score of three on that dimension. The same scores would be assigned on the VAM dimension. That is, teachers would receive a score of one to three on that dimension as well, depending on their rating on the VAM measures. The total points to be awarded to teachers in different cells of the familiar 9-fold table are shown at the top of the next page.

In the system ISR researchers have in mind, annual evaluation ratings would be assigned as in the previous table. That is, three broad classes of teachers would be defined: ineffective teachers (who were rated as ineffective on both performance dimensions), highly effective teachers (who were rated as highly effective on both performance dimensions), and standard teachers (i.e, who were rated neither ineffective nor highly effective on both dimensions). Using three years of accumulated data, but *no* confidence intervals, our simulation showed that under this approach 92.4% of teachers would be rated as "standard", 2% of teachers would be classified as ineffective, and 1% as highly effective using three years of data.[31]

In the ISR approach, annual evaluation ratings are assigned using the three ratings categories just discussed. However, decisions about dismissal and reductions in force would use the *points system* described above. Thus, any teacher whose points total (with three years of data) was 2 would be ineffective and (by law) be dismissed. Reductions in force would occur by ranking teachers (in the pool of affected per

[30] It should be noted, however, that the approach is becoming more widely used in school systems around the country, especially in evaluation systems that rely exclusively on VAM scores to make personnel decisions. However, to our knowledge, the decision approach just described has not been applied to classification decisions using teacher observation scores, nor (to our knowledge) has it been used in systems involving joint classification.

[31] Again, these percentages can be changed. For example, an education authority could classify any teacher who obtained an "ineffective" VAM rating with three years of data to an overall rating of ineffective (increasing the percentage of teachers with an ineffective rating to 6.7%). Or, an education authority could declare that any teacher with an ineffective rating on the VAM *or* observation component would be rated ineffective (leading to about 12% of teachers being classified as ineffective). An education authority also could change its standards for classification (e.g., to + or – 1 sd as opposed to the + or – 1.5 sd). That too would alter the percentages in the cells of the table.

At a Glance: ISR Simulation of Percentage of Teachers Who Would be Placed Into Cells of Joint Rating System Like the One Proposed by MCEE

3 Years of Data (no CI, r = .4, cut point = + or – 1.5sd)

| | Ineffective Observation Rating | Standard Observation Rating | Highly Effective Observation Rating |
|---|---|---|---|
| Ineffective VAM Rating | 2.1% (total points = 2) | 4.6% (total points =3) | 0% (total points =4) |
| Standard VAM Rating | 5.3% (total points = 3) | 77.8% (total points = 4) | 3.4% (total points = 5) |
| Highly Effective VAM Rating | 0% (total points = 3) | 5.7% (total points = 5) | 1% (total points = 6) |

sonnel) according to their points totals, with layoffs proceeding from the lowest ranked employee in the pool upward until the required number of layoffs occurred. Because the ranking system is "coarse" (with 78% of teachers having a score of 4), there is always a strong possibility of tied scores among layoff candidates. Should ties occur, ISR researchers would recommend using the other decision criteria permitted by PA 102 of 2011 (for example, professional contributions) to make a final determination of layoffs.

To be sure, this method of coarse ranking (without statistical confidence regions) has the potential to produce errors of decision making about particular teachers. However, averaging scores across multiple years of data will increase precision somewhat. More importantly, it is well known that:

- ***Over repeated use, personnel selection decisions made from a simple ranking system will always produce higher average performance in an organization than selection via other (non-ranking) methods, as long as the criterion used in rankings have validity.***[32]

- ***Therefore, in the absence of information about measurement precision, ranking is a legally and scientifically defensible approach to making the consequential personnel decisions required by PA 102 of 2011.***

Two final points about the ranking system just described are worth noting. First, the careful reader will note that ISR researchers did *not* assign differential "weights" to scores on the two performance dimensions used in this coarse ranking system. Obviously, ISR researchers are aware that PA 102 calls for school systems to place greater weight on the "student growth" component of annual teacher evaluations in coming years. However, the problem with assigning proportionally greater weight to evidence of student growth in the immediate future is that there is no credible, scientific evidence of the validity of local measures of this performance dimension ( as discussed in Chapter 2). To be sure, locally-developed measures are not entirely lacking in validity, but there is also no *a priori* reason to assign greater decision weight to these measures versus teacher observation scores. In fact, recent research suggests that the best approach to teacher evaluation in situations of fuzzy measurement is to assign *equal* weights to scores on the different performance dimensions used to construct a composite performance index.[33]

Finally, ISR researchers do not believe the points system should be used in annual performance ratings. Instead, ISR researchers think a three category rating system (as shown by red, buff, and green cells of the table) is much more reasonable. The rationale behind having only three ratings is that the thrust of statistical analyses presented in this chapter suggests that teachers in the "buff" colored cells have levels of performance that are, for the most part, statistically indistinguishable. Therefore, a three part classification system for the purposes of annual evaluation, when coupled with consequential standards for tenure and dismissal and a coarse ranking system for reductions in force, would appear (to ISR researchers at least) to be the most warranted approach to making personnel decisions and annual performance ratings in the spirit of section 2(e) of PA 102 of 2011.

[32] See, M.A. Campione, J.L. Outtz, S. Zedeck, F.L. Scmidt, J.F., K.R. Murphy, and R.M Guion, (2001). "The controversy over score banding in personnel selection: Answers to 10 key questions," *Personnel Psychology*, 51(4): 149–185.

[33] See, for example, K. Mihaly, DF McCaffrey, D.O. Staiger and JR Lockwood, *A Composite Estimator of Effective Teaching* (www.metproject.org/MET_Composite_Estimator_of_Teaching_Effectiveness, Gates Foundation, Measures of Effective Teaching Project).

# Chapter 6: Action Steps To Improve Teacher Evaluations in Michigan

Having reviewed data from the pilot of educator effectiveness tools and having explored approaches to improving the evaluation process, this chapter lists a set of action steps that ISR researchers think are needed to build the State of Michigan's capacity to conduct high quality teacher evaluations in light of PA 102 of 2011.

## Improving District Policy and Procedure Manuals

We begin with a mundane but important action step. In Chapter 2 of this report, we noted that many school districts participating in the MCEE pilot of educator effectiveness tools had not yet developed well-crafted and detailed manuals of policy and procedures in the area of teacher evaluation. Yet well-crafted statements of policy and procedure seem warranted if a new system of teacher evaluation practices is to become regularized across all of the schools in a district and transparent to all constituencies involved. Good examples of such manuals exist, and efforts should be made by MDE and professional associations to disseminate such models in order to inform developments in other local education authorities.

Chapter 2 also found that large percentages of teachers in pilot districts wanted more information about the observation tools used in their annual evaluations and did not clearly understand how indicators of student growth were used in their annual evaluations. Districts need to take steps to include teachers in decisions about annual evaluation procedures and educate teachers about procedures in use, especially in areas (like student growth measures) where teachers and administrators have joint responsibility for execution of the district's evaluation policy.

## Improving Classroom Observation Procedures

The data in Chapter 2 of this report suggested that teacher observation procedures in local schools were uneven. Most principals attended four days of base training, but a majority reported that such training did not prepare them to score observations well using the tools assigned to them. As a result, many principals took additional steps, usually through discussions at meetings in their districts. None of this led to strong implementation of classroom observation regimes in the schools. To be sure, principals tended to spread their observation load across the year, and to spread the observation of any given teacher across time—both good sampling procedures. But principals often did not score items on observation tools in the "manner prescribed" by tool vendors and there were low rates of inter-rater reliability.

Data presented in Chapter 3 of this report suggest that the following are central features of good observation practice:

***Training.*** Training in the use of observation tools should consist not only of the 4 days of introductory training provided by vendors at the outset of the pilot, but also additional calibration training designed to improve observation scoring and reduce rater error. ISR research staff engaged in about 6 additional calibration sessions during the pilot (described in Chapter 3), and this improved rates of inter-rater reliability among ISR observers. Calibration training should become a mandatory component of state-provided training in the use of observation tools as a means of developing more accurate scoring of classroom observation tools.

***Fidelity.*** Individuals conducting classroom observations for teacher evaluations should be instructed to use the classroom observation tools in the "manner prescribed" by tool vendors. It will be especially important for principals to score mandatory items on a protocol, for missing item data can affect observation reliability (and perhaps validity).

***Number of observations.*** Data presented in Chapter 3 of this report showed that the reliability of observation scores improves with the number of observations conducted on a teacher. Those data suggested that 3-4 observations per year should be specified as the minimum number of observations per year when a teacher is in an evaluation cycle. Although more observations than that will further increase observation score reliabilities, teachers and principals expressed concerns about the amount of time they were spending on the evaluation process. For this reason, it seems sensible to keep the number of annual observations between 3-4 for most teachers and to further increase reliability by encouraging districts to calculate running averages using up to three years of observation data in annual evaluations.

*Steps to Correct for Rater Error.* Chapter 3 of this report showed that rater error is an important feature of observation scores. As a result, tool vendors and districts should be encouraged to develop procedures for correcting observation scores for rater leniency or severity. A very good way for districts to correct for rater error is to "randomly" assign individuals other than the principal to conduct at least some observations alongside the principal over the course of the school year. This practice should be encouraged by the state. Alternative approaches include using the kinds of statistical adjustments discussed in Chapter 3.

## Improving Measurement of Student Growth

Chapter 2 of this report suggests that one of the least well-implemented aspects of PA 102 of 2011 was the collection of student growth data for use in annual teacher evaluations. The State of Michigan, that chapter showed, does not have a state testing system that can be used easily in annual teacher evaluations, and as Chapter 5 showed, the current system can only be used in the annual evaluations of around 33% of teachers.

In this light, it is not surprising that educators in schools relied mostly on locally-developed tests to fulfill the student growth requirements of PA 102 of 2011. But Chapter 2 of this report showed that there was very little uniformity of measurement in the area of student growth at schools, and the potential at least, that many uses of local tests were not measuring student growth in ways that are consistent with good psychometric practice.

Improvements to implementation of the student growth component of PA 102 of 2011 will require actions by the State and local education agencies together.

***State actions are required.*** Michigan currently does not have the capacity to use value-added measures of teaching effectiveness in its annual teacher evaluation process. Should the State decide to pursue this option, many steps will need to be taken.

***Improved assessment coverage is needed.*** First, the State needs to expand its assessment system to cover more grades and subjects if VAMs are to be used in the teacher evaluation process. If the state wants to pursue the use of VAMs in teacher evaluations, a student assessment system of the sort described in MCEE's final report seems essential.

***Better TSDL data are needed.*** Second, even if a testing system with more grade/subject coverage is implemented, efforts will need to be made to improve the collection of data on teacher-student linkages. Evidence presented in Chapter 4 of this report suggested that the current TSDL (teacher-student data linkage) system is not functioning well, connecting up to 25% of teachers to only a small number of students. This limits the precision (and perhaps validity) of VAM scores calculated from MEAP data. Improvements in this area might be difficult because collection of TSDL data is a complex process shared by local education agencies, which use many different data management tools to interface with the State's SDMS (student data management system) and REP (Registry of Education Personnel). ISR recommends that, first of all, the State undertake a systematic review of the quality of TSDL data and how it is collected and then engage in any required technical upgrades in its own systems or technical assistance to local education agencies that will improve this area of data collection. ISR also recommends that prior to implementing any use of VAM scores in teacher evaluations that the State develop a roster verification process that allows teachers and principals to check the accuracy of data used to estimate VAM scores.

***Making "optional" state assessments available to local districts.*** MCEE's final report described a set of "optional" assessments being developed for use in Michigan's public schools. Data from the pilot research discussed in Chapter 2 of this report suggests that even if such assessments become widely administered in schools, there is no guarantee that they will be used in the teacher evaluation process, as evidenced by the low rates of use in teacher evaluations of the assessment tools provided to schools (at no cost) by the pilot project. ISR researchers suspect that the low incidence of use of pilot tools in teacher evaluations was due—not to resistance by local educators—but rather to a lack of technical knowledge and capacity to use such assessments for teacher evaluation. This suggests that the State of Michigan needs to offer more technical assistance to local schools about the use of well-designed and currently available assessment instruments in teacher evaluation. Many districts, for example, administer standardized achievement tests in grades K-6, but only 20% of teachers reported using data from these tests in their annual evaluations, and many of the assessments used were not good tools for teacher evaluation. In the future, more districts also might begin to use commercially produced "end-of-course" exams (such as those provided by ACT QualityCore). Again, the State needs to offer local districts technical assistance in order for such tests to gain more widespread acceptance and use in annual teacher evaluations.

***Developing Better Local Measures of Student Growth.*** Since responsible implementation of a value-added component of teacher evaluations seems several years away, and since local educators have expressed a strong preference for using locally-developed tests as measures of student growth in teacher evaluations, it seems very likely that in the near future, the student growth component of PA 102 of 2011 will depend crucially on good use of local assessments. Chapter 2 of this report described some potential shortcomings in the use of local assessments in teacher evaluations conducted in pilot schools. One way to improve local practice would be for the state to provide better technical assistance in the use of local assessments in the teacher evaluation process.

## Assignment of Effectiveness Ratings to Teachers

Chapter 6 of this report described some of the issues associated with classifying teachers into the effectiveness ratings defined in section 2(e) of PA 102 of 2011. The chapter argued that any personnel evaluation system needs to set standards of performance, measure performance, and understand the degree of statistical (un)certainty of these measures when it makes personnel decisions. Chapter 2 of this report showed that current practices in pilot districts departed from this process to some extent. Districts *did* have standards (although they were apparently not uniform across districts). Districts also had measures, but apart from implementation of 4 observation protocols, measures of student growth and professional growth varied widely from district-to-district, and no districts attempted to assess the degree of statistical uncertainty in these measures or use such information in assigning effectiveness ratings to teachers. Further, districts gave widely varying weights to different performance criteria in their formulae for assigning effectiveness ratings to teachers. As a result, the new teacher evaluation process being implemented as a result of PA 102 of 2011 is *not* leading to uniform classification of teachers into the effectiveness ratings required under the law.

***Setting Performance Standards.*** The state should convene a panel of educators and researchers and engage in a standards-setting exercise that sets recommended levels of performance for a teacher to be rated above "ineffective" in an annual evaluation. This involves determining what scores are to be obtained on the state-recommended observation protocols, what scores could be used from value-added modeling, and what processes could be used to set uniform performance standards on commercially-developed and locally-developed assessments commonly used in schools.

***Vendor Assessment of Measurement Precision.*** ISR recommends that *any* observation vendor or VAM vendor having a contract with the state provide the state and local education agencies with standard errors of measurement for use in setting confidence intervals around observation and VAM scores. This is common practice in educational measurement and is already a piece of information provided by the pilot's VAM vendors. It is *not*, however, something that is routinely provided by observation vendors—despite it being a common practice in the education measurement community among both test makers and researchers. Observation vendors have the technical capacity to develop psychometrically sound measures and provide SEMs to clients. State contracts should insist on the provision of this information.

***Local Assessment of Measurement Precision.*** It will be more difficult for local education authorities to quantify measurement precision of their locally-developed measures—especially locally-developed measures of student growth. It also seems unwise to demand that local entities develop procedures to calculate precision. In most cases, local districts will lack both the technical expertise and the capacity to develop highly sophisticated measures of precision.

***Even in the absence of information about measurement precision, PA 102 of 2011 will continue to be in force, and districts must continue not only to assign***

***teachers to effectiveness ratings annually, but also use evaluation information for consequential personnel decisions. ISR researchers recommend that, until a more rigorous system of measurement is put into place by the State, districts use a three-category classification system in annual evaluations and a simple ranking approach when making dismissal and layoff decisions.*** In this process: (a) the state will work to promote consensus and provide guidelines about standards for placing teachers into three effectiveness ratings (ineffective, standard, and highly effective); (b) an initial rating of teachers will be developed from two rating criteria—student growth and classroom observations; (c) any teachers who is rated as ineffective on both criteria three years in a row can be dismissed under the requirements of PA 102 of 2011; (d) reductions in force can be handled through a simple ranking procedure in which a teacher's overall score is a simple sum of his or her ratings (on a scale of 1-3) on classroom observations and student growth, and reductions in force will be enforced simply by always choosing the teacher among the list of potential teachers subject to reduction in force that has the lowest ranking according to this formula. Ties can be handled by including additional data—including data on professional responsibilities and professional growth. Importantly, while a simple ranking system of this sort will produce some errors in decision making, it is well known that over the long run, it works to increase the mean performance level in organizations. In the absence of information about precision of measurement, it is therefore a rational procedure that is widely used by organizations and recommended by experts in the field of personnel psychology.

## Timing of Improvement Steps

The list of action steps just presented represents an enormous effort. As a result, ISR recommends that a new teacher evaluation system be rolled out over a period of at least three years.

### Observation Tool Rollout

***Begin Rolling out Observation Tools.*** The rollout of the teacher observation component of Michigan's new teacher evaluation system seems like an immediate target of action. Chapter 2 showed that most principals thought the observation tools they piloted were better than what they had used in the past, and both teachers and principals found them to provide reasonably accurate depictions of classroom practice. Moreover, the vendors that ISR worked with had well-developed training procedures and technical infrastructure that could be implemented at scale.

***Recommended Tools.*** On the basis of pilot data, ISR is prepared to strongly recommend the use of the Danielson FFT observation protocol as a state-approved tool. FFT has good measurement properties, a well-functioning technical system for capturing and reporting observation data to schools, and is well-positioned to provide calibration training. ISR also recommends 5D and the Thoughtful Classroom (TC) tools for state adoption, although with slightly less enthusiasm than FFT. The TC tool had the highest rated technical platform and, for the most part, ISR researchers found it to be a well-functioning instrument from a psychometric standpoint. The main problem with TC from ISR's perspective was the high percentage of items (other than the "four corners") that were rated with low frequency by principals. If this instrument is to be used in schools, ISR recommends that the "four corners" items be rated on *all* observation occasions and that scores used in final teacher evaluations be based on teacher scores on these items. The 5D instrument has many positive features—including a focus on intellectually demanding work in classrooms—but it is a long instrument and during the pilot, its technical systems were not rated as highly by principals or used as much by them as were the other vendors' technical systems. If the state uses 5D as a vendor, it should explore development of a "short form" of the instrument and investigate the status of its technical systems. Finally, ISR researchers want to express important reservations about the Marzano instrument. To begin, this is an extremely complex instrument, and in the pilot project, the instrument was filled out with very low rates of agreement about: (a) when to score items; and (b) the scores to be assigned on items that were scored. Of the four instruments piloted, this instrument was the hardest to work with from a psychometric standpoint, and as a result, ISR researchers were unwilling to develop well-established IRT models using data from the instrument.

***Training by Observation Tool Vendors.*** Once the state decides which vendors it wants to work with it will need to release requests for cost proposals. However, prior to taking this step, ISR recommends convening observation vendors and cognizant state officials to discuss the array of services the state wants to purchase and the timing of any rollout of training. Interviews with observation tool vendors conducted by ISR researchers suggest that it might be difficult for any vendor to complete statewide training of all clients in a one-year period, and that both initial and validation training might be pursued in partnership with local organizations.

## VAM Rollout

The rollout of any system of value-added measurement of teaching effectiveness will require numerous steps.

***Assessment Development.*** Obviously, use of VAM scores in teacher evaluations requires an expanded state testing system. The rollout of this system will take time, probably proceeding along the timeline discussed in MCEE's *Final Recommendations* document.

***TSDL Development.*** During the time period that these new state assessments are under development, the State would be wise to take immediate action to better structure its data systems for value-added modeling. An immediate first step would be to improve the socio-technical systems involved in capturing and verifying TSDL data. As discussed in Chapter 4 of this report, up to 25% of teachers in the VAM pilot were associated with as few as 7 students using the current data system, and this compromised the technical quality of VAM estimates. As a result, there is a need to investigate the current data collection system thoroughly, find gaps in data processing, and correct any shortcomings through technical assistance to local districts.

***Verification System.*** Even if a well-functioning TSDL data system is developed, ISR recommends implementation of a roster verification system in Michigan. This will provide an important chance for those affected by the data system to verify that data used in potentially "high stakes" decisions are accurate from their point of view and can be another step where errors are corrected and important data are added to the TSDL data.

***Learning about VAMs.*** Value-added modeling is a complex endeavor, and VAM vendors use many different approaches to estimate teaching effectiveness. During the course of the pilot, ISR researchers became concerned that many key education constituencies might not fully understand the technical details of value-added modeling and the required policy choices that need to be made by key decision makers. ISR researchers therefore recommend that cognizant state officials convene meetings with the VAM vendors to learn more about approaches to value added measurement, to gain advice about how to proceed, and to begin to put in place any contractual specifications the state will require. Such convenings also should include technical experts and state educators, whose views on the issues should be taken into account.

***ISR researchers have no strong preferences about choice of a VAM vendor.*** Each vendor is technically competent and can provide strong services to Michigan. Choice of a vendor, ISR assumes, will come down to a competitive bidding process that selects the one vendor who can provide services the State lists in its request for proposals at a competitive price.

## Standards Setting and Classification

The education community in Michigan needs to develop a more uniform understanding about standards of teaching effectiveness to be used in teacher evaluations. In addition to implementing the standards setting process described earlier in this chapter, ISR recommends two additional steps.

***CEPI should receive all teacher observation data arising from the use of state-approved teacher observation tools (for research purposes only).*** Such data can be obtained from vendor databases, and transmittal of data to the state should be a part of any contract with observation tool vendors. Using such data the state (or a qualified contractor) can explore appropriate psychometric models to apply to these data and how to quantify the precision of estimates of teaching practice derived from such models. This work should be undertaken in conjunction with (and inform) the statewide standards setting process discussed above. At issue in this research are: (a) the score values that will be considered as cutoffs for classifying teachers into the different effectiveness ratings mandated by section 2(e) of PA 102 of 2011; and (b) the relative precision of decision making that is advisable (i.e, the confidence intervals desired by decision makers in the classification process).

***The State of Michigan also should contract with a vendor to estimate VAM scores on a test basis using existing state assessment (and perhaps other) data.*** At a minimum, research with state assessment data can evaluate the effectiveness of efforts to improve state TSDL data by examining the number of student-teacher linkages available in existing state data before and after any TSDL improvement projects. Such research also could examine whether and how any such improvements affect the technical quality of VAM score estimates.

### Costs

In a separate report, ISR is providing cost estimates for implementing different configurations of a statewide system to support high quality teacher evaluation. That report will be disseminated to the public shortly after release of the present report.

UNIVERSITY OF
MICHIGAN

To Contact the First Author:

Brian Rowan
Burke A. Hinsdale Collegiate Professor in Education
Research Professor, Institute for Social Research
Professor of Sociology

Email: [illegible]
Phone: 734/647-3648
Mail: Education and Well Being Program
ISR Survey Research Center
The University of Michigan
2354 Perry Building
330 Packard Street
Ann Arbor, MI 48104-1248